Research Report

Investigating the Effectiveness of a Synthetic Linking Function on Small Sample Equating

Sooyeon Kim

Alina A. von Davier

Shelby Haberman

Research & Development

August 2007
RR-07-37

Investigating the Effectiveness of a Synthetic Linking Function on Small Sample Equating

Sooyeon Kim, Alina A. von Davier, and Shelby Haberman

ETS, Princeton, NJ

August 2007

Abstract

The synthetic function, which is a weighted average of the identity (the trivial linking function for forms that are known to be completely parallel) and a traditional equating method, has been proposed as an alternative for performing linking with very small samples (Kim, von Davier, & Haberman, 2006). The purpose of the present study was to investigate the benefits of the synthetic function using various real data sets gathered from different administrations of tests from a licensure testing program. We investigated the chained linear, Tucker, Levine, and mean equating methods, along with the identity and the synthetic functions with small samples (N = 19 to 70). Neither the identity nor the synthetic functions worked as well as did other linear equating methods, because test forms differed markedly in difficulty. The synthetic function cannot be used as a solution or methodological fix to a problem that is caused by poor data collection design.

Key words: Linking bias, equating error, synthetic function, identity, small sample

Acknowledgments

This paper was presented at the annual meeting of the National Council on Measurement in Education (NCME) held April 9 to 13, 2007, in Chicago, IL.

We would like to thank Samuel Livingston, Neil Dorans, Michael Walker, and Dan Eignor for their many helpful comments on earlier drafts of this paper. The authors also gratefully acknowledge the editorial assistance of Kim Fryer.

Table of Contents

Page

Introduction ... 1

Small-Sample Equating ... 1

An Alternative Method for Small Sample Equating: The Synthetic Function ... 2

Purpose ... 4

Methodology ... 4

Design ... 4

Equating/Linking Functions Used in This Study ... 5

Criterion Function ... 6

Procedure ... 6

Deviance Measure ... 7

Study 1 ... 7

Data ... 7

Criterion Function ... 9

Results ... 11

Study 2 ... 26

Data ... 26

Criterion Function ... 27

Results ... 27

Discussion ... 35

References ... 39

List of Tables

Page

Table 1. Descriptive Statistics of Each Sample in Study 1 8

Table 2. Summary Statistics for the Observed Distributions of *X*, *V* in *P* and *Y*, *V* in *Q*: Study 1 9

Table 3. RMSD Between the Criterion Function and Sample-Based Linking Functions (Identity, Mean, Chained Linear, and Synthetic) Across the Entire Score Region: Study 1 12

Table 4. RMSD Between the Criterion Function and Sample-Based Linking Functions (Tucker, Levine, and Synthetic) Across the Entire Score Region: Study 1................ 13

Table 5. The 90% Confidence Interval for RMSDs Calculated for the 100 Bootstrap Samples Across the Entire Score Region (Mean, Chained, Linear, and Synthetic): Study 1 ... 14

Table 6. The 90% Confidence Interval for RMSDs Calculated for the 100 Bootstrap Samples Across the Entire Score Region (Tucker, Levine, and Synthetic): Study 1................ 15

Table 7. Descriptive Statistics of Each Sample in Study 2 26

Table 8. Summary Statistics for the Observed Distributions of *X*, *V* in *P* and *Y*, *V* in *Q*: Study 2 27

Table 9. RMSD Between the Criterion Function and Sample-Based Linking Functions (Identity, Mean, Chained Linear, and Synthetic) Across the Entire Score Region: Study 2 29

Table 10. RMSD Between the Criterion Function and Sample-Based Linking Functions (Tucker, Levine, and Synthetic) Across the Entire Score Region: Study 2................ 29

Table 11. The 90% Confidence Interval for RMSDs Calculated for the 100 Bootstrap Samples (Mean, Chained Linear, and Synthetic) Across the Entire Score Region: Study 2 34

Table 12. The 90% Confidence Interval for RMSDs Calculated for the 100 Bootstrap Samples Across the Entire Score Region (Tucker, Levine, and Synthetic): Study 2................ 34

List of Figures

Page

Figure 1. Difference plot, chained linear versus chained equipercentile, and frequency distribution of form *X* scores in Total Group P in Study 1. 10

Figure 2. September 2003 administration (*N* = 19). 16

Figure 3. September 2004 administration (*N* = 23). 16

Figure 4. January 2005 administration (*N* = 23). 17

Figure 5. March 2004 administration (*N* = 25). 17

Figure 6. January 2004 administration (*N* = 26). 18

Figure 7. September 2005 administration (*N* = 28). 18

Figure 8. August 2005 administration (*N* = 30). 19

Figure 9. November 2003 administration (*N* = 36). 19

Figure 10. March 2005 administration (*N* = 42) 20

Figure 11. January 2006 administration (*N* = 42). 20

Figure 12. April 2004 administration (*N* = 46). 21

Figure 13. November 2004 administration (*N* = 46). 21

Figure 14. November 2005 administration (*N* = 46) 22

Figure 15. April 2005 administration (*N* = 49) 22

Figure 16. April 2003 administration (*N* = 53) 23

Figure 17. June 2003 administration (*N* = 59) 23

Figure 18. June 2005 administration (*N* = 65). 24

Figure 19. June 2004 administration (*N* = 69) 24

Figure 20. Difference plot, chained linear versus chained equipercentile, and frequency distribution of form X scores in total group P in Study 2. 28

Figure 21. August 2005 administration (*N* = 31) 30

Figure 22. September 2004 administration (*N* = 35). 30

Figure 23. January 2005 administration (*N* = 56). 31

Figure 24. April 2004 administration (*N* = 64). 31

Figure 25. November 2005 administration (*N* = 66). 32

Figure 26. April 2005 administration (*N* = 70). 32

Introduction

Test equating is a statistical method that makes scores from different test forms interchangeable by adjusting for differences in difficulty among forms that are built to the same specifications. As are other statistical procedures, the equating of test scores is subject to sampling effects such as error or bias. If the sample is large and representative, the equating relationship in the sample may accurately represent the equating relationship in the population. The smaller the sample, though, the more likely it is that the equating function computed for that particular sample differs substantially from that of the population. Additionally, sampling error may affect the extent to which the sample represents the population from which it was drawn and, as a result, can influence the quality of the equating. In practice, the use of a small sample can be expected to have more influence on equating when equating samples are not representative.

Small-Sample Equating

Estimated equating relationships contain estimation error. Random equating error, which is typically indexed by the standard error of equating (SEE), is present whenever samples are used to estimate equating relationships in populations (Kolen & Brennan, 2004). Sample size has a direct effect on the SEE, which is the measure of the statistical accuracy of estimated equating functions. A few empirical studies have examined the use of equating with small samples with respect to equating error or bias (Livingston, 19 93; Parshall, Du Bose Houghton, & Kromrey, 1995; Skaggs, 2005).

Parshall et al. (1995) examined the effects of sample size (e.g., 15, 20, 50, and 100) on the stability and bias of linear equating with two parallel forms based on the NEAT design. Their results suggested trivial levels of equating bias, even with small samples, but substantial increases in SEEs as sample size decreased. The sampling error was smallest in the proximity of the mean raw score of the test, and the error increased monotonically (but not linearly) as a function of the deviation of scores from the mean. This finding implies that the SEE associated with differences in sample size becomes more pronounced for scores at greatest distances from the mean raw score for the test.

In a similar manner, Skaggs (2005) used an equivalent-groups design to study the equating of the passing score on a certification test, using samples ranging from 25 to 200 observations. As expected, the SEE became smaller as sample size increased; however, equating bias changed little as a function of sample size. Even the sample that included 200 observations evinced substantial equating error on at least part of the raw score scale, yielding a significant percentage of examinees whose pass/fail designations were incorrectly specified. Skaggs also found that, with samples as small as 25, not equating is likely to do less harm to examinees than is some form of equating, because equating with such small samples may produce a degree of equating error that could exceed the total equating error variance, at least when using linear equating methods.

Some experts believe that use of an identity function or no equating is preferable to nontrivial equating with extremely small samples, as the large random equating error associated with very small samples negates the benefits of equating (Harris, 1993; Kolen & Brennan, 2004). In many testing programs (e.g., in programs with a consistently low volume of examinees), it may not be feasible to obtain large samples. Nevertheless, these programs need to report, in a timely manner, comparable scores over different administrations or test forms. Many practitioners may confront this dilemma in situations where form construction has not been well-structured. Accordingly, research on equating with insufficient data is necessary to obtain some guidelines that are useful in practice. In our previous study (Kim, von Davier, & Haberman, 2006), we proposed an alternative method (called a synthetic function) that can provide certain empirical benefits with very small equating samples, including reduction of equating error. The explanation for this method is presented in the next section.

An Alternative Method for Small Sample Equating: The Synthetic Function

An alternative to equating with small samples is the use of the identity function, which is appropriate for linking when forms are completely parallel. The use of the identity function is equivalent to not conducting any equating at all. Formally, the identity function is

$$ID_Y(x) = x, \tag{1}$$

where x is a raw score of the new form X that is placed on the raw scale of the old form Y in a direct linear manner with a slope of 1 and intercept of 0. The random equating error is zero for identity linking, because the equated (or linked) scores are obtained through a deterministic procedure. However, the use of the identity function can increase systematic error (i.e., bias). In general, the identify function is appropriate when test specifications are well-defined and two forms are nearly parallel in both difficulty and content.

The synthetic function is essentially a compromise between using the sample equating function and using no equating (through use of an identity function) by combining them using a specified weight system:

$$\text{Synthetic}_y(x) = w \times CL_y(x) + (1\text{-}w) \times ID_y(x), \tag{2}$$

where w is a weight between 0 and 1; x is raw score in form X; CL_y is the chained linear function (which can be replaced by other types of equating functions); and ID_y is the identity function.

As discussed previously, equating with small samples may lead to large sampling error reflected in the SEE. Given that the SEE for the identity linking is zero, the SEE for the synthetic linking can be substantially reduced as compared to the chained linear function used here. To demonstrate the extent to which the equating errors can be reduced when using the synthetic function, Equations 3 and 4 are presented:

$$\begin{aligned} Var(Syn_y(x)) &= w^2 Var(CL_y(x)) + (1-w)^2 Var(ID_y(x)) + 2w(1-w)Cov(CL_y(x), ID_y(x)) \\ &= w^2 Var(CL_y(x)). \end{aligned} \tag{3}$$

Hence,

$$SEE(Syn_y(x)) = w \times SEE(CL_y(x)). \tag{4}$$

From Equation 4, we see that when the same weight is given to the identity and chained linear functions, equating error is reduced by one half. Similarly, Equation 5 shows that the bias of the synthetic function that is mostly introduced by use of the identity function can be reduced under the assumption that the chained linear equating function is not biased or much less biased when compared to the identity function.

$$\mu\,(\text{Synthetic}_y[x]) = w \times (\text{Mean}\,[CL_y(x)]) + (1\text{-}w) \times (\text{Mean}\,[ID_y(x)]). \tag{5}$$

In our previous research (Kim et al., 2006), we examined the linking bias and error among the identity, chained linear, and synthetic functions using data sets from two different types of national assessments: one with a highly reliable external anchor and one with a less reliable internal anchor. In both assessments, the synthetic function and even the identity function outperformed the chained linear method based on very small samples. The chained linear function showed the greatest amount of linking error, although its bias was relatively small. Root mean squared error for the synthetic function was smaller than that of the identity function when sample sizes were greater than 100. The synthetic function exhibited lower linking error at the expense of a small amount of bias. Because the test forms used in our previous research were well-designed and almost parallel, using the identity function was likely to do less harm to examinees than using conventional equating in this case.

Purpose

The purpose of the present study was to extend our previous work by investigating the benefits of the synthetic function using various real data sets over different administrations of tests from a licensure program. In the previous study, small samples were randomly drawn from large samples, and the two test forms were almost parallel. In the present study, the two forms were clearly nonparallel, and very few tests were administered. The use of the identity function might be inappropriate when substantial differences exist between forms or between the shapes of their respective score distributions. The findings derived from various data sets, which were collected at different testing administrations, would be informative in assessing the effectiveness of the synthetic function in practice. The two weight systems (e.g., equal and unequal) were used to synthesize the sample equating function with the identity function.

Methodology

Design

A nonequivalent groups anchor test (NEAT) design is often used with small samples in practice. In the NEAT design, there are two test forms, *X* and *Y,* to be equated, and a target population, *T*, for which the equating is done. These two operational tests (*X* and *Y*) are given to two samples of examinees from different test populations or administrations (usually denoted by the populations *P* and *Q*). Accordingly, the two test scores, *X* and *Y*, are each only observed either on *P* or on *Q*, but not on both. An anchor test is given to samples from both *P* and *Q*. The anchor test score can be either a part of both *X* and *Y* (the internal anchor case) or a separate score (the external anchor case). Tests on two different types of subject matter for a licensure program, mostly composed of low-volume samples, were selected for this study. The data sets from these tests were collected using the NEAT equating design with common internal anchors.

Equating/Linking Functions Used in This Study

We examined four linear equating methods (equating form *X* to form *Y*), along with the identity and synthetic functions, to investigate the benefits of the synthetic function as an alternative to equating with small samples. The equating methods were: (a) mean, (b) chained linear, (c) Tucker, and (d) Levine observed-score methods. Because the samples of interest were too small to ensure the adequacy of equipercentile equating results (Harris, 1993; Kolen & Brennan, 2004), only linear equating functions were considered in this study. Many observed-score equating methods were based on the linear equating function. All these functions and their (untestable) assumptions have been described in detail elsewhere (Kolen & Brennan, 2004; von Davier & Kong, 2005).

The synthetic function creates a compromise between the identity function (no equating) and linear linking by combining them using a specific weight system. One issue related to the use of the synthetic function concerns the manner in which the two functions are averaged using a certain weight system. No literature or other criteria provide definitive guidance regarding the weight systems to be used when averaging the two different equating/linking functions. The best approach in this case is to re-evaluate the data collection design, beginning with the design of the test itself. However, more often the weight system for the two functions is investigated after collecting data. In that case, as explained in the previous study (Kim et al., 2006), the a priori information about the test forms and the anchor (e.g., sample size, tests and anchor reliability, test specifications, and test variability over time) and information about the two populations of examinees can be considered as guidelines in deciding on specific weights.

As mentioned previously, in this study both equal and unequal weights were applied based upon test characteristics. Although weight systems for the two functions are flexible, a simple scheme to illustrate the use of the synthetic function weights the two functions equally. A more thorough analysis for choosing appropriate weights is an interesting issue for further research that will be presented elsewhere. A formula that transforms the ordinary weight system into the symmetric weight system also was described in detail elsewhere (Kim et al., 2006).

Criterion Function

Two different tests were administered to groups of less than 100 (range = 19–70). Defining a criterion for the types of low-volume data sets that were collected in the NEAT design is difficult. As an alternative, data sets accumulated over several administrations were treated as the pseudo-populations *P* and *Q* for examinees who took test form *X* or *Y*, respectively. An equating function derived from those pseudo-populations was then used as a criterion in this study. In the NEAT design, information would only be available on *X* in *P* and on *Y* in *Q*. Therefore, the criterion in a NEAT design will be admissible when the correlations between total scores and anchors are high enough (e.g., over .90). Both linear and nonlinear equating methods were used in this study to determine a criterion function. The criterion function (i.e., raw-to-equated raw conversion) for each test will be explained in detail later.

Procedure

As a first step, the new form *X* was equated to the reference form *Y* with the total accumulated samples, using both linear and nonlinear equating methods. Then these equated *X* scores were used as the equating criterion. As the next step, we obtained the form *X* scores equated to form *Y* with a single-administration sample using identity, mean, chained linear, Levine, and Tucker methods. The identity function was combined with the chained linear, Tucker, and Levine methods, respectively, to create various synthetic functions in each administration. Based upon test characteristics, the two weight systems were employed mainly to combine any linear equating function with the identity function. These were (a) equal weight (.5) for both functions and (b) more weight for the actual equating function (.7) than for the identity function (.3).

The same examinees who took form *Y* and were used to define a criterion were used as the reference sample in each administration. The operational sample who took form *X* in each administration was used as a new form sample. Accordingly, the new form *X* sample was much smaller than the reference form *Y* sample, which is the usual case in practical equating situations. The equating functions/results observed in each administration were compared with the equating criterion derived from the total groups in each single administration with respect to the following deviance measure.

Deviance Measure

To evaluate equating/linking results, the differences among the conversions were quantified using the root mean squared difference (RMSD, see Equation 6) across both the entire range of raw scores and the cut-score region in each administration.[1] The data sets used in this study were from licensure tests; thus equating accuracy was much more important at the passing score region than at any other score points. In practice, one way to assess the small sample equating result is to examine its impact on examinees' pass/fail designations. For that reason, RMSD was also examined solely for the cut-off score region.

The average equating difference is given by,

$$RMSD = \sqrt{\sum_{i=0}^{I} w_i \left[\hat{e}_{yi}\left(x_i\right) - e_{yi}\left(x_i\right) \right]^2} \qquad (6)$$

where *i* represents each score point, $\hat{e}_{yi}\left(x_i\right)$ represents the equated scores of an equating method at raw score *x*, $e_{yi}\left(x_i\right)$ is the criterion equating function, and w_i is the relative proportion of examinees at each score point.

As auxiliary information, the 90% confidence intervals (CIs) of RMSDs were calculated for each equating/linking function using a resampling technique. Using the SAS PROC SURVEYSELECT procedure that randomly selects units with replacement, 100 bootstrap samples (i.e., 100 replications) composed of the same number of examinees were randomly selected from *P* (those who took form *X*) in each administration; then, form *X* scores were equated to form *Y* for those 100 samples. The RMSDs were calculated for those 100 samples to obtain the 90% CI for all the equating/linking functions.

Study 1

Data

The descriptive information for the data sets used in Study 1 is presented in Table 1. As shown, there were 18 individual administrations of form *X* being employed. The data sets for form *X* were composed of samples with relatively low volumes, ranging from 19 to 69. Each data set consisted of the raw sample frequencies of scores for two nonparallel, 60-item tests[2] with 20

Table 1

Descriptive Statistics of Each Sample in Study 1

Test form	Administration		*N*	Total		Anchor		r_{xv}
				M	*SD*	*M*	*SD*	
Form *X*	September	2003	19	36.16	8.06	14.00	2.96	.94
	September	2004	23	36.04	7.67	13.61	2.55	.92
	January	2005	23	35.17	7.77	13.43	3.06	.92
	March	2004	25	38.16	6.69	14.16	2.88	.83
	January	2004	26	33.85	7.25	12.27	2.71	.84
	September	2005	28	36.75	9.60	13.82	3.37	.93
	August	2005	30	38.03	8.20	14.00	2.79	.93
	November	2003	36	33.78	7.56	12.22	3.36	.87
	March	2005	42	38.83	7.86	14.40	2.97	.87
	January	2006	42	37.02	8.54	13.71	3.44	.86
	April	2004	46	35.89	8.59	13.61	3.45	.89
	November	2004	46	35.20	8.38	13.15	3.16	.90
	November	2005	46	36.11	8.33	13.24	3.14	.91
	April	2005	49	38.47	8.71	13.86	3.40	.92
	April	2003	53	36.53	9.32	13.25	3.47	.90
	June	2003	59	37.85	8.16	14.08	3.01	.89
	June	2005	65	37.08	7.73	13.58	3.25	.88
	June	2004	69	36.80	9.03	13.58	3.54	.89
Form *Y*	Total		426	40.38	7.38	12.86	3.04	.89

Note. r_{xv} indicates correlation between total and anchor scores.

internal anchor items given to two samples (called *P* and *Q*) from a national population of examinees. Sample *P* comprised examinees who took the new test form *X*, and sample *Q* comprised those who took the reference form *Y*. The total number of examinees who took form *X* from April 2003 to January 2006 was 678,[3] and the total number who took form *Y* from March 1998 to March 2003 was 426. Descriptive statistics for these groups are summarized in Table 2.

As shown in Table 2, the mean of the anchor test *V* was 13.72 (± 0.12) in total group *P*, and 12.86 (± 0.15) in total group *Q*, where 0.12 and 0.15 were the standard errors of the mean. Thus, total group *Q* was less proficient than total group *P*, as measured by *V*. In terms of effect sizes, the difference between the two means (0.86) was approximately 28% of the averaged standard deviation of 3.12. This magnitude indicated a fairly large difference between the two groups for this type of testing program. However, psychometric properties (e.g., standard error of measurement, reliability, correlation between total score and anchor) for the two forms were fairly similar. The internal-consistency reliabilities of anchors were much lower than those of total tests, because internal anchors consisted of fewer items than did total tests. The correlations between tests and anchors was reasonably high ($r = .89$).

Table 2

Summary Statistics for the Observed Distributions of X, V in P and Y, V in Q: Study 1

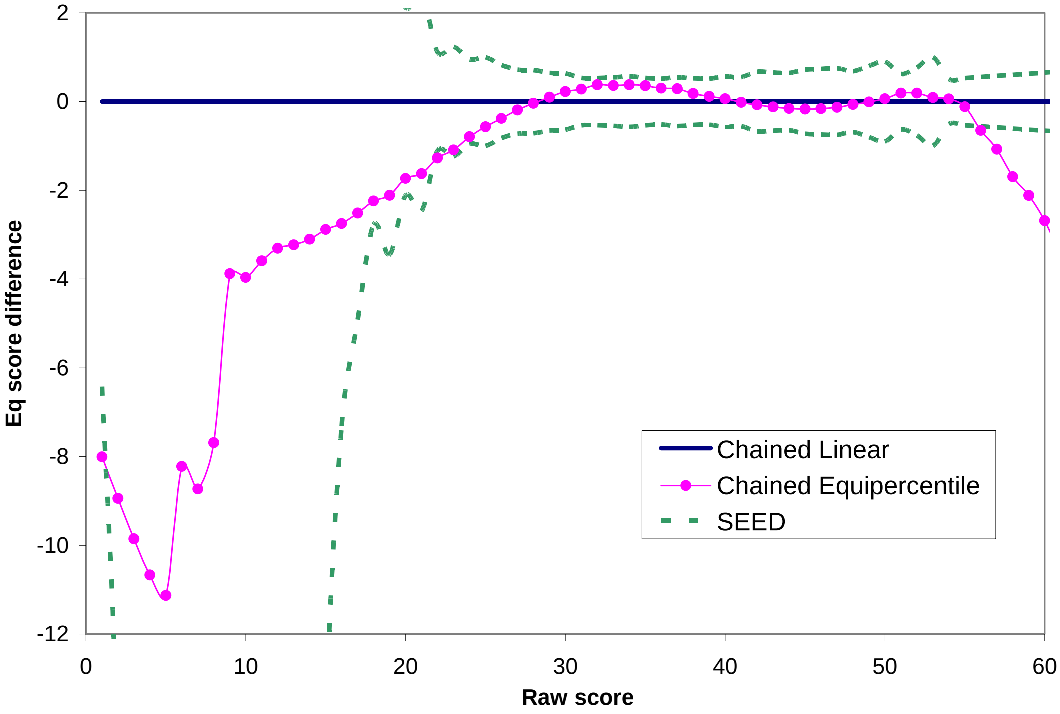
Distributions	*N*	μ	σ	SEM	Reliability	ρ
X_P	678	37.19	8.26	3.14	.85	.89
V_P		13.72	3.20	1.76	.69	
Y_Q	426	40.38	7.38	3.25	.80	.89
V_Q		12.86	3.04	1.84	.63	

Note. SEM = Standard error of measurement. ρ = Correlation between total score and anchor. *P* = Accumulated from April 2003 to January 2006 administrations. *Q* = Accumulated from March 1998 to March 2003 administrations.

Criterion Function

Equating was conducted with a total of 678 examinees for form *X* and 426 examinees for form *Y*. For each raw score on form *X*, the equivalent raw scores on form *Y* were determined using the chained linear, Tucker, Levine, and chained equipercentile methods.

Figure 1 presents equated raw score differences between the chained linear and chained equipercentile methods, along with the frequency distribution of form *X* scores in the total group, *P*. The differences between the two equating functions were very large for the score points from 0 to 25 and from 57 to 60. Among the 678 examinees who took the new form *X*, the minimum



Raw Score X	Frequency
0-3	0
4-6	0
7-9	0
10-12	0
13-15	1
16-18	4
19-21	12
22-24	23
25-27	47
28-30	66
31-33	86
34-36	75
37-39	87
40-42	80
43-45	85
46-48	52
49-51	38
52-54	13
55-57	9
58-60	0
Total	678

***Figure 1.* Difference plot, chained linear versus chained equipercentile, and frequency distribution of form *X* scores in Total Group *P* in Study 1.**

observed score was 14, and only 8% (*N* = 51) of examinees received scores lower than 26. No data are available for raw score points from 58 to 60. This means that the differences of the two functions observed at the low and very high score points may be artificial, due to a lack of data. For raw score points from 25 to 57, which included most examinees, differences between the two functions were less than the *difference that matters* (Dorans & Feigenbaum, 1994), defined as half of a score point. In addition, the differences between those two functions were trivial at the cut-score region (27 to 36). Differences between the linear and nonlinear functions were assumed to be negligible for this case. In addition, at less than 1,000, the samples were not large enough to support the nonlinear function here. Based on those observations, the raw-to-equated raw conversion derived from the chained linear function was considered as a criterion and was compared with equating functions derived from operational samples in each administration.

Results

Using data sets from 18 test administrations, form *X* was equated to form *Y* using various linear equating methods along with the identity and synthetic functions. Tables 3 and 4 present RMSDs across the entire range of raw scores and the cut-off score region (Raw Scores 27 to 36). The 90% confidence intervals of RMSDs derived from the resampling technique are presented in Tables 5 and 6. In addition, Figures 2 to 19 present plots of the differences between the sample equating function (e.g., chained linear, identity, and synthetic) and the criterion (horizontal line in each figure) over the entire raw score range in the 18 administrations.

The RMSD for the identity function (5.32) was much larger than that for the other equating functions in all administrations. As shown in Table 2, the mean of form *Y* (M_{YQ} = 40.38), taken by less able examinees, was much higher than that of form *X* (M_{XP} = 37.19), taken by more able examinees. This means that the reference form *Y* was much easier than the new form *X*. Because Forms *X* and *Y* differed markedly in difficulty, the use of the identity function yielded enough bias to pose a problem in a certification testing program. Consequently, the synthetic functions partially based on the identity function also yielded quite large RMSDs compared to those for the traditional linear equating methods.

Among the four equating methods (i.e., mean, chained linear, Tucker, and Levine), no method always performed better than the others in dealing with small samples in test equating. As expected, equating results from chained linear, Tucker, and Levine methods were fairly

Table 3

RMSD Between the Criterion Function and Sample-Based Linking Functions (Identity, Mean, Chained Linear, and Synthetic) Across the Entire Score Region: Study 1

Test administration	*N*	Identity		Mean		Chained linear		Synthetic (.5CH+.5ID)		Synthetic (.7CH+.3ID)	
1	19	5.32	(3.38)	1.93	(0.94)	1.64	(1.11)	1.76	(1.07)	0.42	(0.19)
2	23	5.32	(3.38)	1.18	(0.47)	1.27	(0.83)	2.20	(1.18)	1.21	(0.39)
3	23	5.32	(3.38)	1.43	(0.63)	1.22	(0.69)	2.02	(1.32)	0.74	(0.51)
4	25	5.32	(3.38)	0.49	(0.20)	0.86	(0.35)	2.74	(1.86)	1.75	(1.25)
5	26	5.32	(3.38)	0.56	(0.36)	0.58	(0.20)	2.85	(1.75)	1.90	(1.12)
6	28	5.32	(3.38)	1.00	(0.36)	0.96	(0.67)	2.24	(1.29)	1.14	(0.50)
7	30	5.32	(3.38)	0.50	(0.16)	0.95	(0.39)	2.56	(1.43)	1.61	(0.71)
8	36	5.32	(3.38)	0.52	(0.31)	1.13	(0.47)	2.80	(1.93)	1.86	(1.34)
9	42	5.32	(3.38)	0.52	(0.13)	0.24	(0.16)	2.53	(1.57)	1.44	(0.87)
10	42	5.32	(3.38)	0.52	(0.13)	0.34	(0.07)	2.60	(1.70)	1.53	(1.04)
11	46	5.32	(3.38)	1.15	(0.45)	1.03	(0.48)	2.17	(1.44)	0.95	(0.67)
12	46	5.32	(3.38)	0.80	(0.22)	0.48	(0.34)	2.37	(1.48)	1.23	(0.74)
13	46	5.32	(3.38)	0.49	(0.24)	0.28	(0.06)	2.69	(1.67)	1.68	(1.01)
14	49	5.32	(3.38)	1.02	(0.73)	0.89	(0.56)	3.08	(1.95)	2.20	(1.39)
15	53	5.32	(3.38)	0.67	(0.47)	0.64	(0.24)	2.88	(1.76)	1.94	(1.14)
16	59	5.32	(3.38)	0.62	(0.11)	0.46	(0.32)	2.43	(1.49)	1.34	(0.76)
17	65	5.32	(3.38)	0.55	(0.35)	0.70	(0.41)	2.84	(1.90)	1.86	(1.30)
18	69	5.32	(3.38)	0.50	(0.17)	0.09	(0.03)	2.62	(1.68)	1.56	(1.01)

Note. Cut-score ranges in parentheses.

Table 4

RMSD Between the Criterion Function and Sample-Based Linking Functions (Tucker, Levine, and Synthetic) Across the Entire Score Region: Study 1

Test administration	*N*	Tucker		Synthetic (.5Tucker +.5ID)		Synthetic (.7Tucker +.3ID)		Levine		Synthetic (.5Levine +.5ID)		Synthetic (.7Levine +.3ID)	
1	19	1.49	(1.01)	1.84	(1.13)	0.52	(0.26)	1.71	(1.16)	1.72	(1.05)	0.38	(0.16)
2	23	1.01	(0.69)	2.24	(1.27)	1.16	(0.49)	1.37	(0.90)	2.18	(1.14)	1.21	(0.35)
3	23	1.11	(0.62)	2.08	(1.35)	0.82	(0.56)	1.28	(0.72)	2.00	(1.31)	0.71	(0.49)
4	25	1.08	(0.69)	3.01	(2.04)	2.12	(1.50)	0.83	(0.21)	2.62	(1.78)	1.59	(1.13)
5	26	0.24	(0.11)	2.72	(1.71)	1.71	(1.06)	0.78	(0.25)	2.91	(1.76)	2.00	(1.14)
6	28	0.96	(0.60)	2.34	(1.32)	1.31	(0.55)	0.99	(0.70)	2.19	(1.27)	1.06	(0.47)
7	30	0.85	(0.25)	2.66	(1.52)	1.71	(0.83)	0.99	(0.46)	2.51	(1.39)	1.55	(0.66)
8	36	0.97	(0.32)	2.69	(1.84)	1.70	(1.22)	1.24	(0.55)	2.85	(1.97)	1.94	(1.40)
9	42	0.38	(0.15)	2.78	(1.73)	1.79	(1.09)	0.44	(0.31)	2.40	(1.49)	1.27	(0.77)
10	42	0.22	(0.16)	2.73	(1.75)	1.71	(1.11)	0.55	(0.09)	2.54	(1.69)	1.45	(1.01)
11	46	0.72	(0.39)	2.28	(1.47)	1.08	(0.72)	1.20	(0.53)	2.13	(1.43)	0.91	(0.65)
12	46	0.43	(0.31)	2.41	(1.49)	1.29	(0.76)	0.51	(0.36)	2.35	(1.47)	1.20	(0.73)
13	46	0.38	(0.09)	2.74	(1.69)	1.74	(1.03)	0.24	(0.04)	2.67	(1.66)	1.65	(1.00)
14	49	1.12	(0.63)	3.17	(1.98)	2.34	(1.43)	0.80	(0.52)	3.03	(1.93)	2.13	(1.37)
15	53	0.85	(0.26)	2.93	(1.76)	2.03	(1.14)	0.53	(0.23)	2.85	(1.77)	1.89	(1.15)
16	59	0.35	(0.13)	2.60	(1.59)	1.56	(0.90)	0.57	(0.41)	2.35	(1.44)	1.22	(0.69)
17	65	0.72	(0.50)	2.93	(1.94)	1.99	(1.36)	0.74	(0.38)	2.79	(1.88)	1.81	(1.28)
18	69	0.26	(0.09)	2.72	(1.70)	1.71	(1.05)	0.28	(0.05)	2.58	(1.68)	1.50	(1.01)

Note. Cut-score ranges in parentheses.

Table 5

The 90% Confidence Interval for RMSDs Calculated for the 100 Bootstrap Samples Across the Entire Score Region (Mean, Chained, Linear, and Synthetic): Study 1

Test administration	*N*	Mean	Chained linear	Synthetic (.5CH+.5ID)	Synthetic (.7CH+.3ID)
1	19	(0.95, 2.87)	(0.76, 2.72)	(1.37, 2.28)	(0.14, 1.21)
2	23	(0.56, 2.17)	(0.54, 2.36)	(1.81, 2.57)	(0.57, 1.74)
3	23	(0.55, 2.30)	(0.51, 2.62)	(1.63, 2.65)	(0.47, 1.61)
4	25	(0.49, 1.79)	(0.25, 3.03)	(1.82, 3.48)	(0.52, 2.86)
5	26	(0.49, 1.95)	(0.17, 2.02)	(2.16, 3.49)	(1.06, 2.78)
6	28	(0.49, 2.03)	(0.49, 1.92)	(1.75, 2.66)	(0.53, 1.77)
7	30	(0.49, 1.19)	(0.42, 1.82)	(2.27, 3.10)	(1.16, 2.32)
8	36	(0.49, 1.47)	(0.36, 2.56)	(2.18, 3.40)	(1.26, 2,78)
9	42	(0.49, 1.51)	(0.23, 1.69)	(1.95, 3.10)	(0.96, 2.24)
10	42	(0.49, 1.57)	(0.17, 1.74)	(2.08, 3.15)	(0.92, 2.31)
11	46	(0.54, 2.16)	(0.26, 2.09)	(1.66, 2.58)	(0.38, 1.53)
12	46	(0.52, 1.60)	(0.26, 1.44)	(1.95, 2.80)	(0.68, 1.84)
13	46	(0.49, 1.19)	(0.21, 1.47)	(2.27, 3.20)	(1.13, 2.36)
14	49	(0.52, 1.65)	(0.35, 1.69)	(2.66, 3.46)	(1.64, 2.73)
15	53	(0.49, 1.47)	(0.31, 1.55)	(2.46, 3.30)	(1.43, 2.52)
16	59	(0.49, 1.33)	(0.17, 1.37)	(2.03, 2.89)	(0.95, 1.96)
17	65	(0.49, 1.21)	(0.32, 1.58)	(2.52, 3.28)	(1.41, 2.50)
18	69	(0.49, 1.14)	(0.19, 1.20)	(2.23, 3.09)	(1.04, 2.21)

Table 6

The 90% Confidence Interval for RMSDs Calculated for the 100 Bootstrap Samples Across the Entire Score Region (Tucker, Levine, and Synthetic): Study 1

Test administration	*N*	Tucker	Synthetic (.5Tucker +.5ID)	Synthetic (.7Tucker +.3ID)	Levine	Synthetic (.5Levine +.5ID)	Synthetic (.7Levine +.3ID)
1	19	(0.65, 2.55)	(1.44, 2.37)	(0.17, 1.25)	(0.82, 2.80)	(1.32, 2.27)	(0.19, 1.20)
2	23	(0.33, 2.01)	(1.79, 2.58)	(0.56, 1.69)	(0.66, 2.51)	(1.79, 2.56)	(0.56, 1.83)
3	23	(0.46, 2.67)	(1.65, 2.66)	(0.43, 1.69)	(0.56, 2.59)	(1.61, 2.65)	(0.46, 1.61)
4	25	(0.32, 3.25)	(2.11, 3.89)	(0.88, 3.42)	(0.25, 3.10)	(1.65, 3.35)	(0.36, 2.81)
5	26	(0.19, 2.62)	(1.97, 3.36)	(1.00, 2.62)	(0.24, 2.09)	(2.28, 3.54)	(1.20, 2.86)
6	28	(0.60, 1.60)	(1.87, 2.77)	(0.72, 1.86)	(0.50, 2.10)	(1.66, 2.63)	(0.44, 1.74)
7	30	(0.42, 1.68)	(2.32, 3.19)	(1.24, 2.42)	(0.37, 1.89)	(2.22; 3.04)	(1.12, 2.26)
8	36	(0.32, 2.67)	(2.14, 3.25)	(1.08, 2.49)	(0.43, 2.80)	(2.23, 3.48)	(1.26, 2.92)
9	42	(0.18, 1.61)	(2.26, 3.38)	(1.23, 2.62)	(0.25, 1.96)	(1.85, 3.00)	(0.89, 2.10)
10	42	(0.16, 1.62)	(2.16, 3.28)	(1.02, 2.54)	(0.25, 1.88)	(1.99, 3.08)	(0.75, 2.23)
11	46	(0.15, 1.63)	(1.89, 2.66)	(0.56, 1.62)	(0.33, 2.29)	(1.53, 2.55)	(0.34, 1.52)
12	46	(0.20, 1.42)	(2.00, 2.85)	(0.75, 1.89)	(0.26, 1.45)	(1.95, 2.79)	(0.69, 1.81)
13	46	(0.19, 1.49)	(2.33, 3.19)	(1.20, 2.36)	(0.22, 1.57)	(2.25, 3.17)	(1.09, 2.36)
14	49	(0.51, 1.84)	(2.76, 3.52)	(1.80, 2.85)	(0.26, 1.67)	(2.57, 3.45)	(1.55, 2.72)
15	53	(0.42, 1.65)	(2.53, 3.36)	(1.47, 2.64)	(0.29, 1.51)	(2.42, 3.27)	(1.38, 2.48)
16	59	(0.11, 1.07)	(2.21, 3.04)	(1.17, 2.15)	(0.18, 1.55)	(1.93, 2.83)	(0.84, 1.87)
17	65	(0.33, 1.50)	(2.63, 3.37)	(1.58, 2.60)	(0.28, 1.63)	(2.45, 3.26)	(1.32, 2.46)
18	69	(0.16, 1.05)	(2.34, 3.14)	(1.19, 2.28)	(0.12, 1.34)	(2.16, 3.10)	(0.94, 2.25)

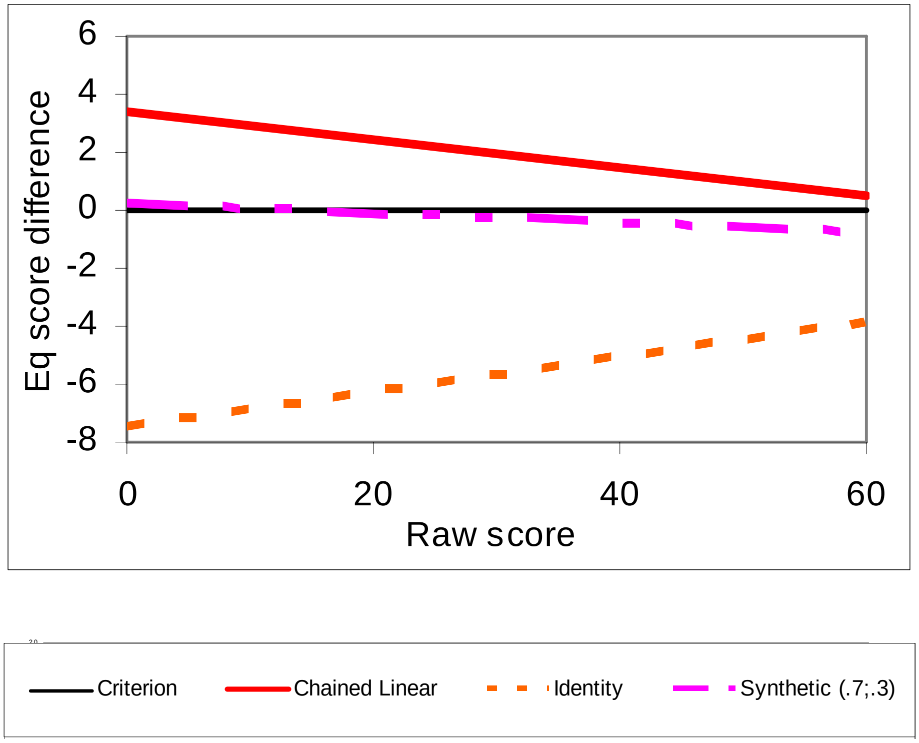


***Figure 2*. September 2003 administration (*N* = 19).**

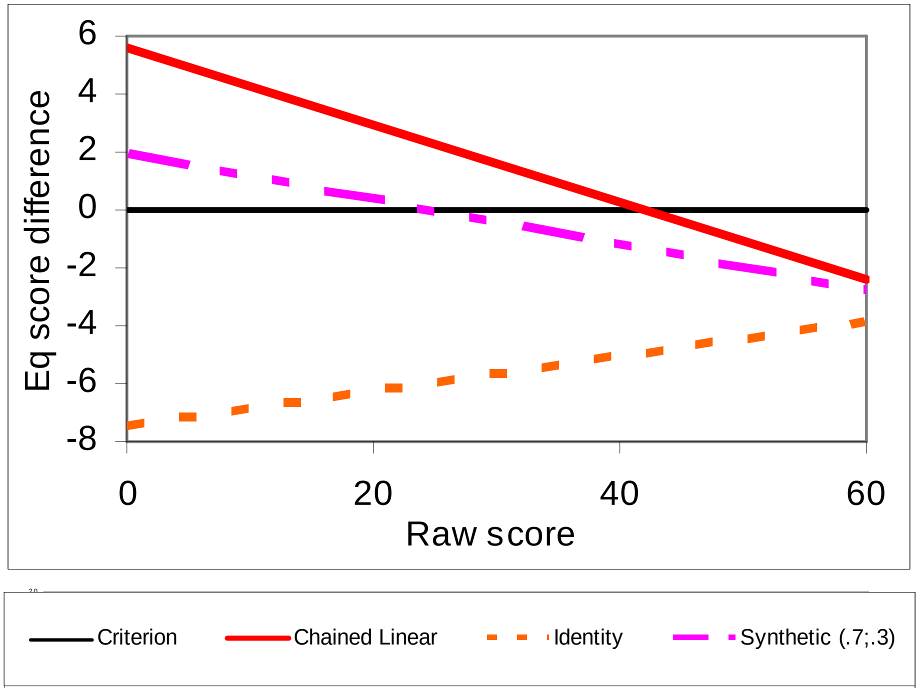


***Figure 3*. September 2004 administration (*N* = 23).**

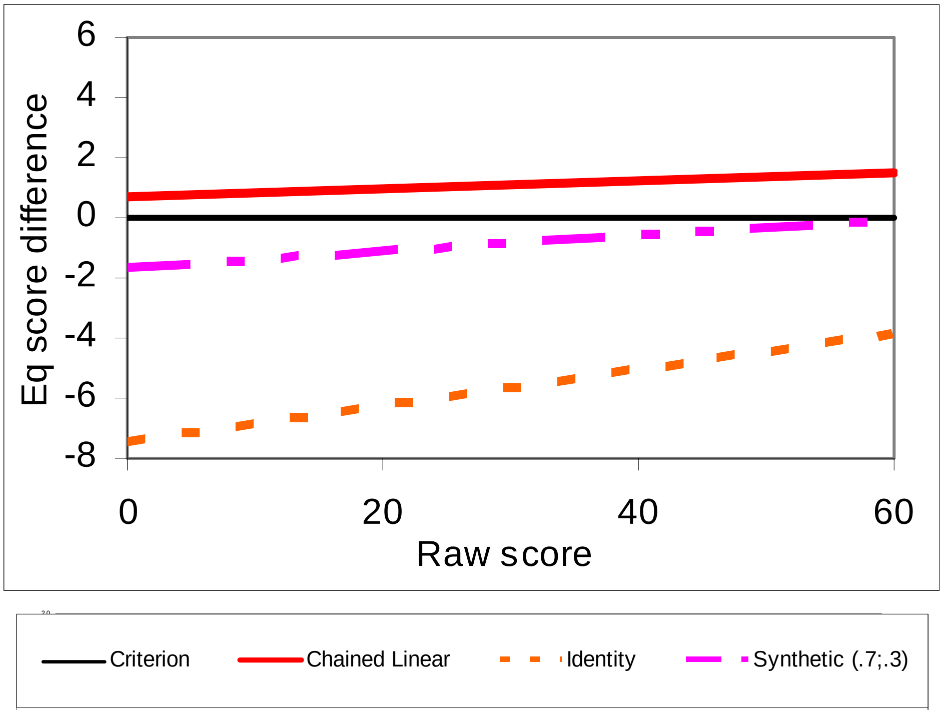


***Figure 4*. January 2005 administration (*N* = 23).**

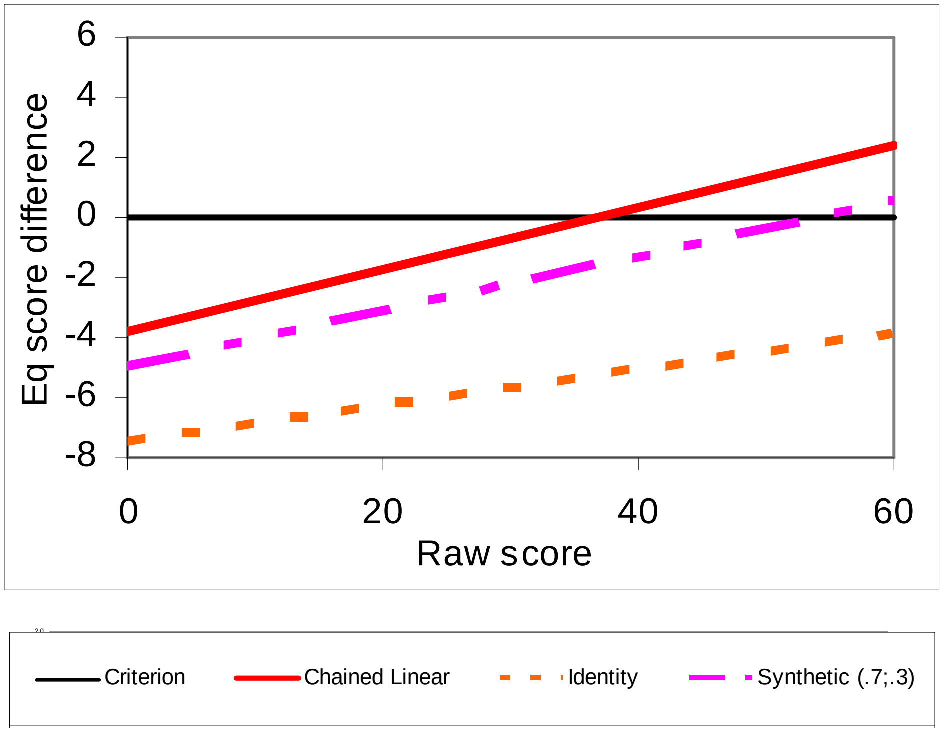


***Figure 5*. March 2004 administration (*N* = 25).**

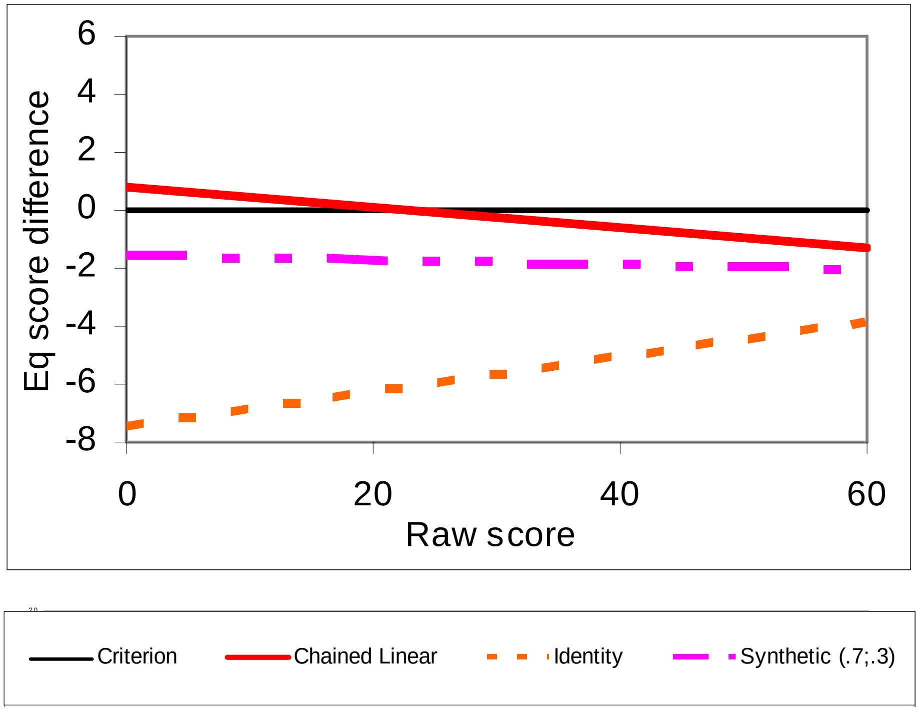


***Figure 6*. January 2004 administration (*N* = 26).**

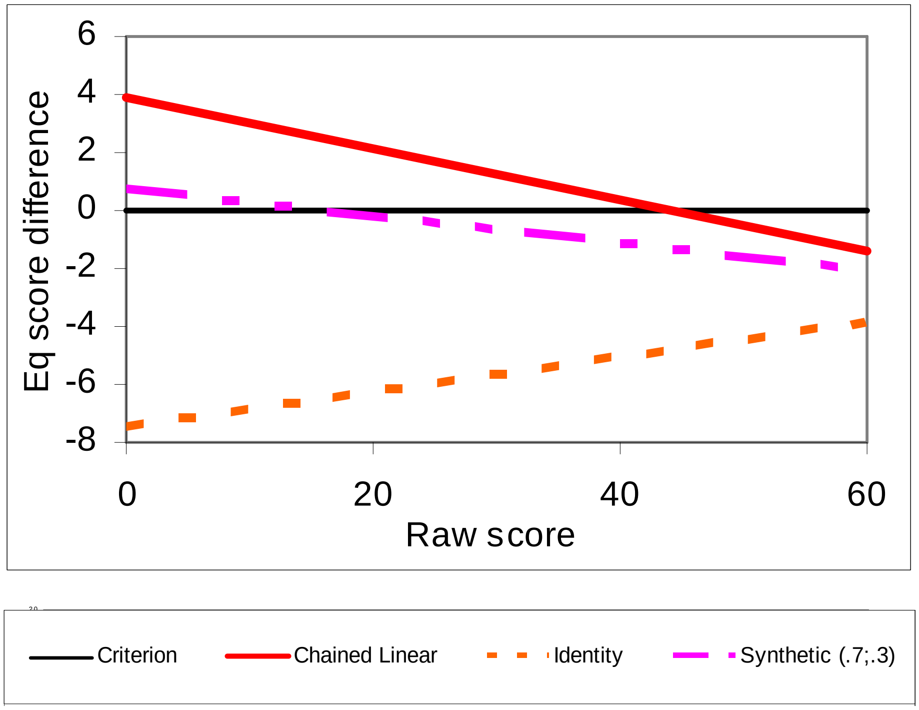


***Figure 7*. September 2005 administration (*N* = 28).**

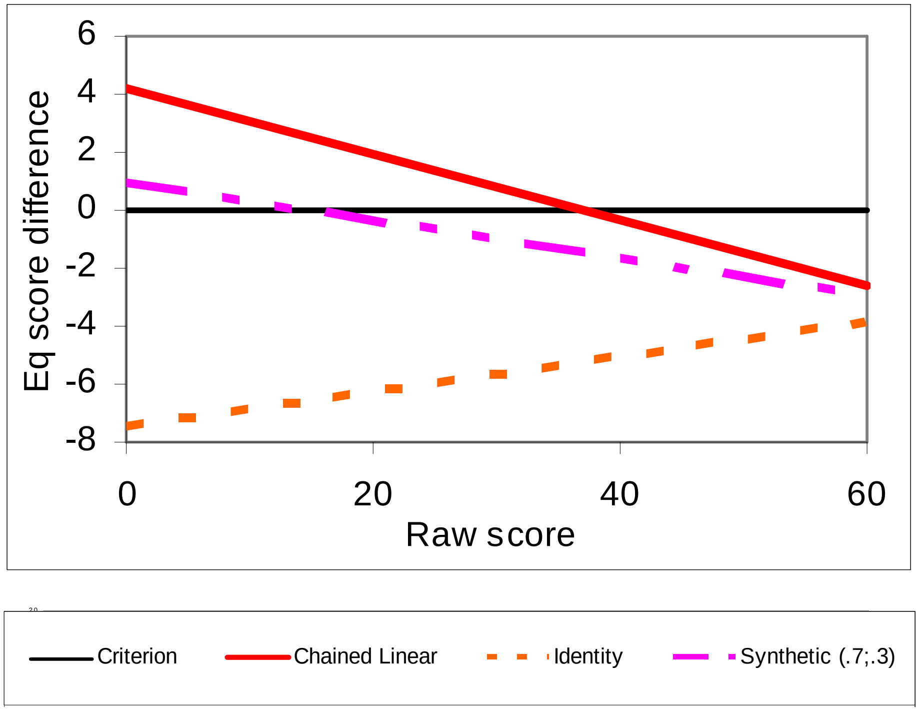


***Figure 8*. August 2005 administration (*N* = 30).**

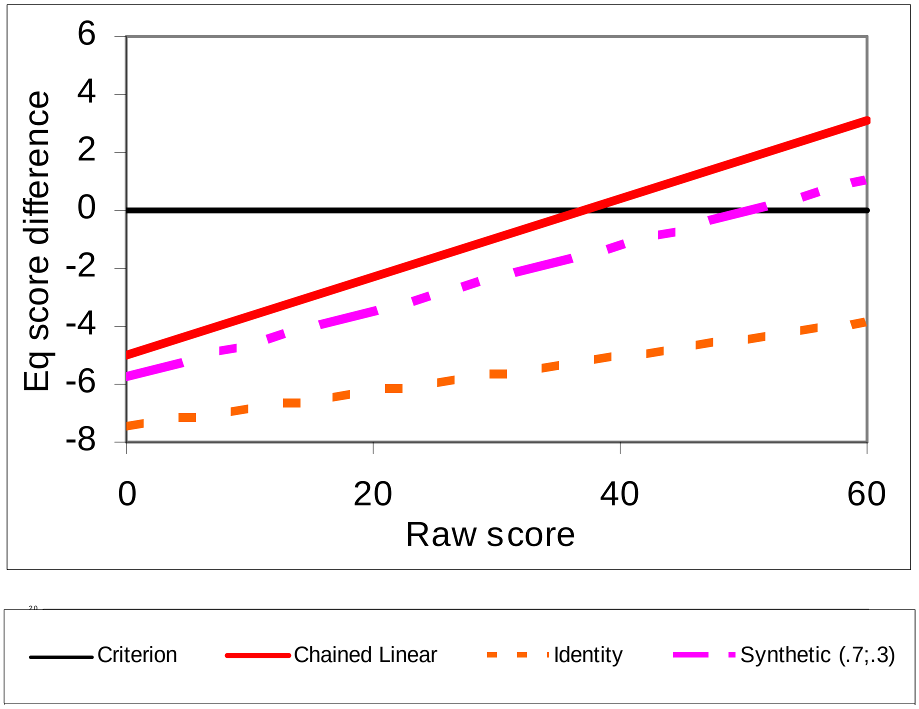


***Figure 9*. November 2003 administration (*N* = 36).**

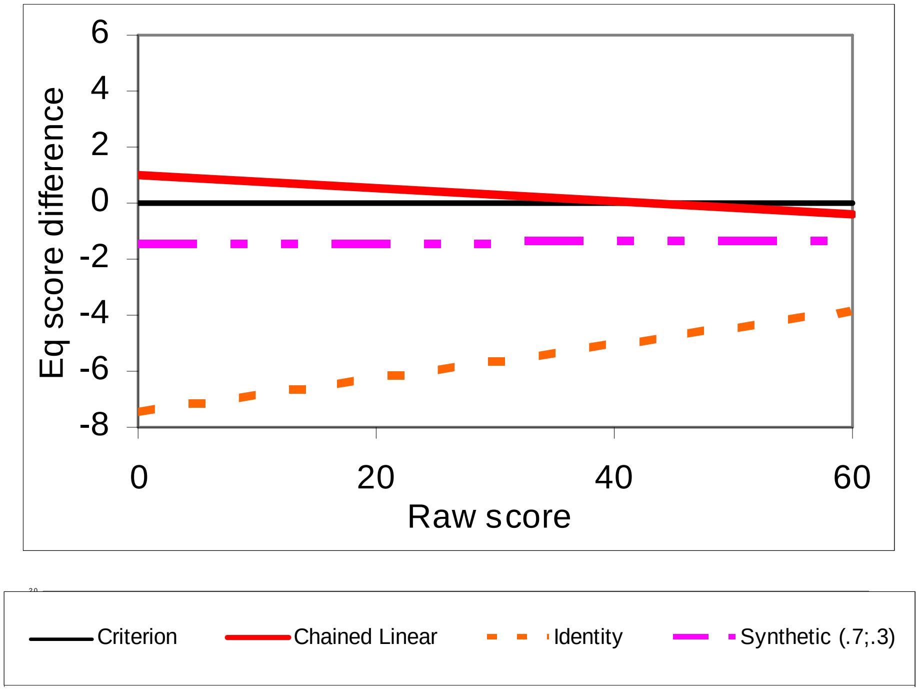


***Figure 10*. March 2005 administration (*N* = 42).**

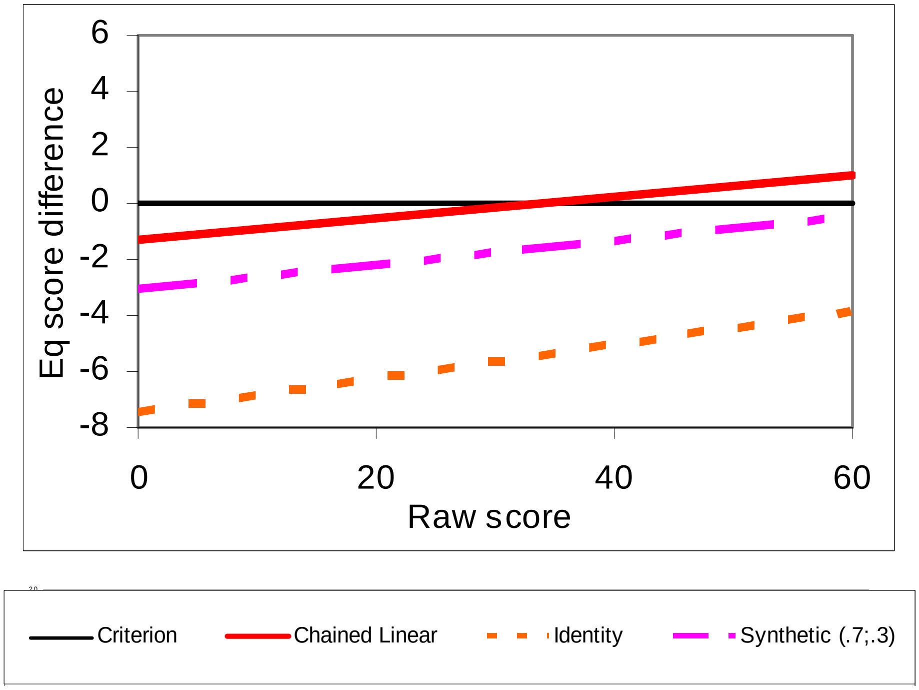


***Figure 11*. January 2006 administration (*N* = 42).**

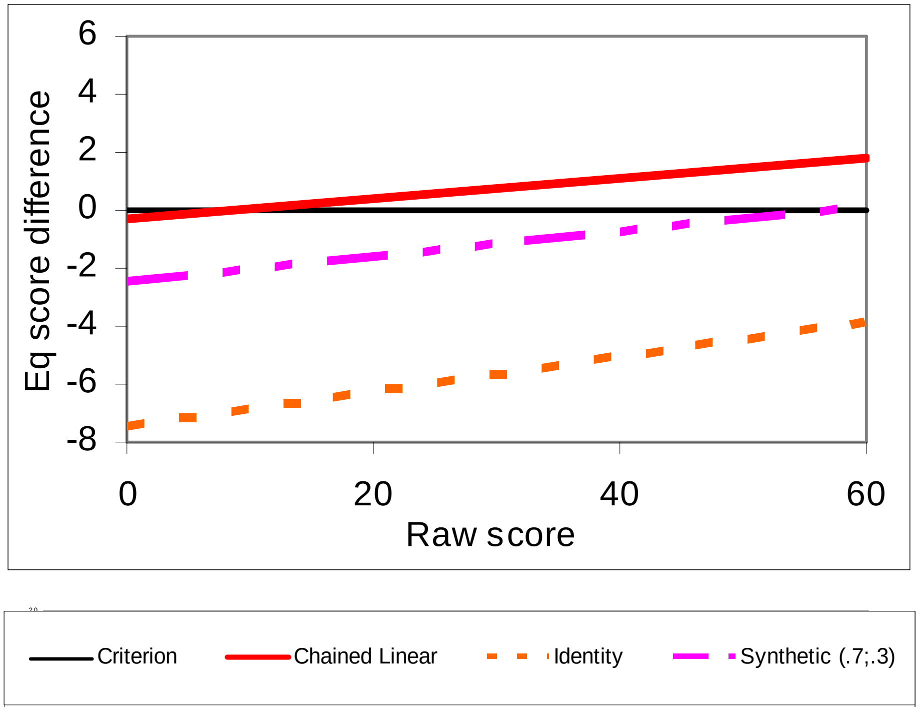


***Figure 12*. April 2004 administration (*N* = 46).**

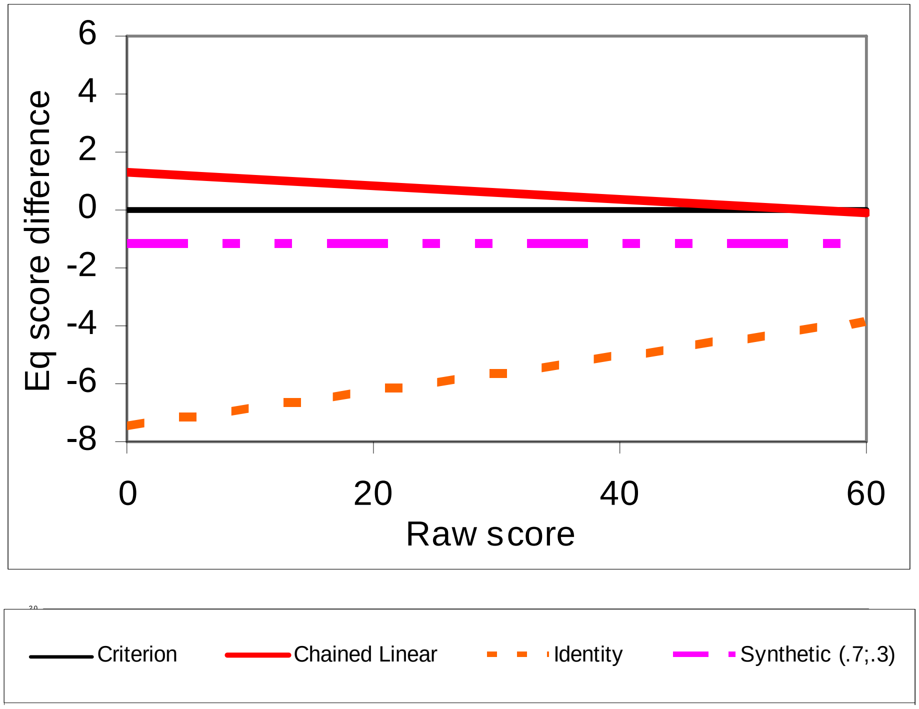


***Figure 13*. November 2004 administration (*N* = 46).**

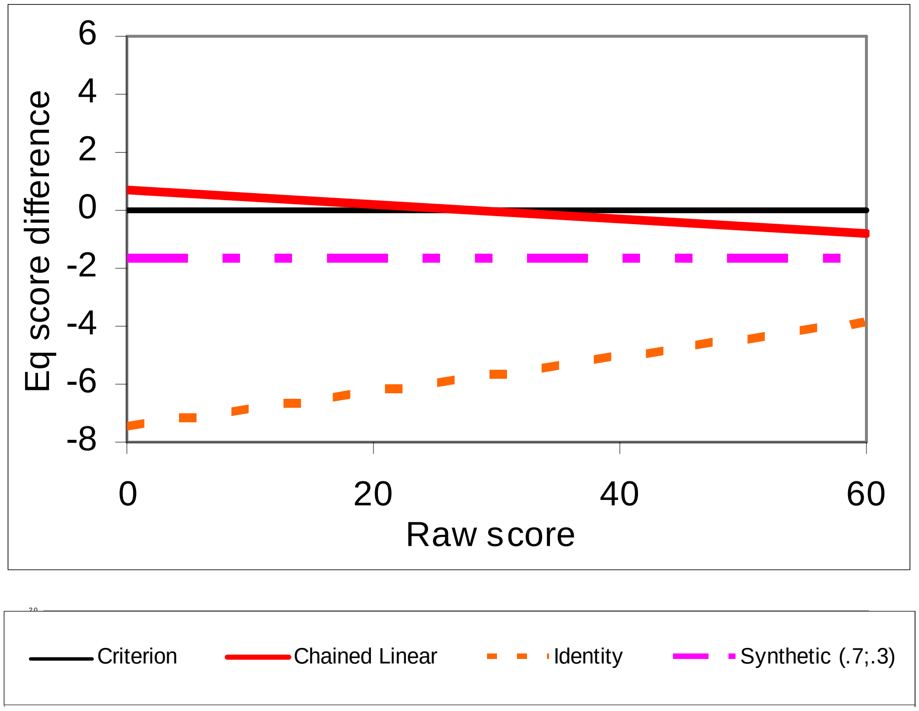


***Figure 14*. November 2005 administration (*N* = 46).**

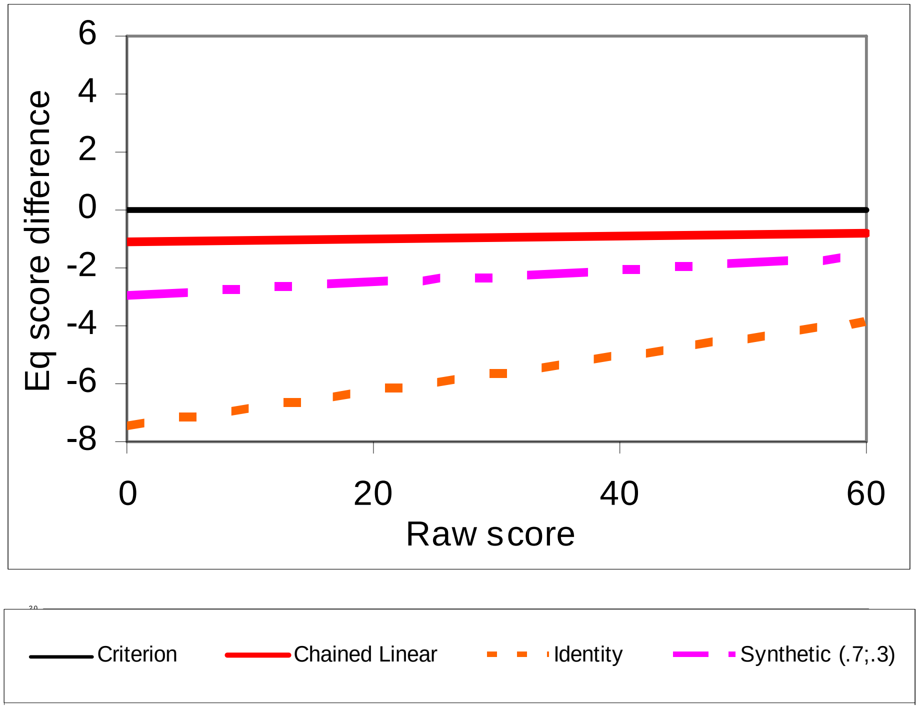


***Figure 15*. April 2005 administration (*N* = 49).**

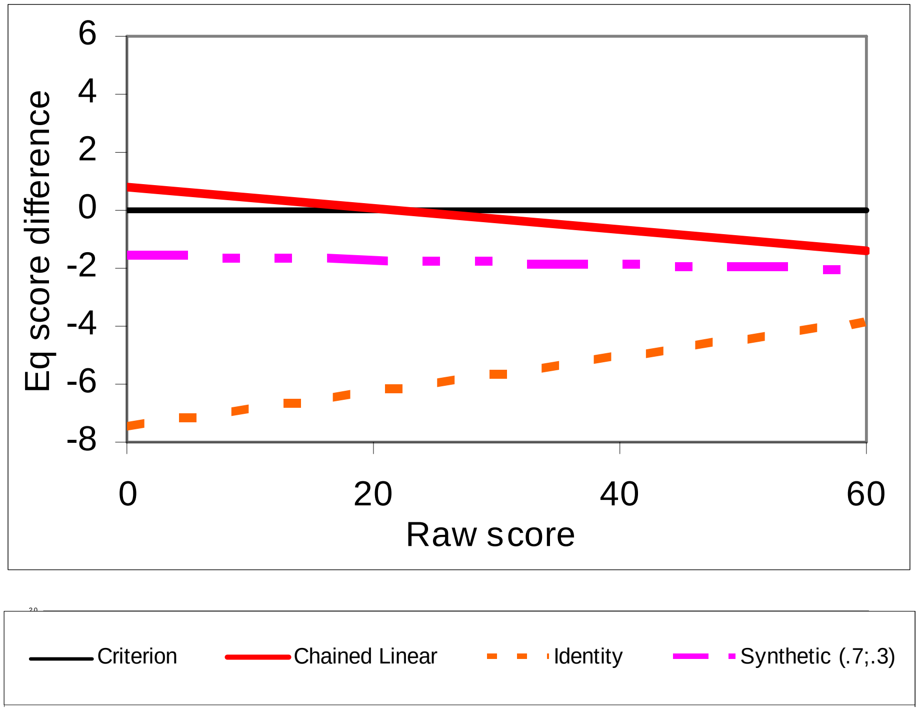


Figure 16. **April 2003 administration (*N* = 53).**

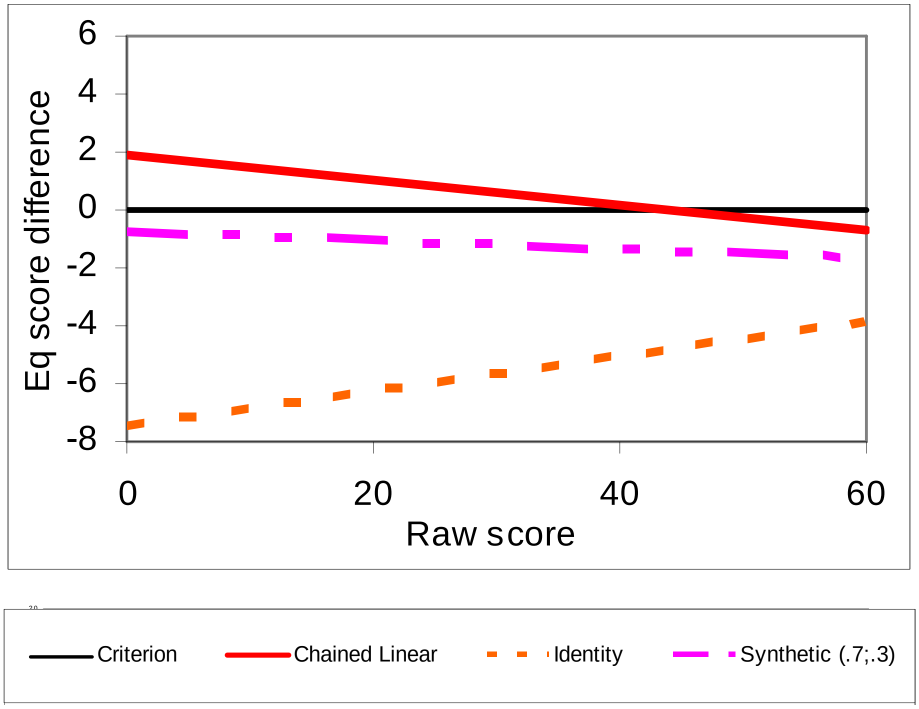


Figure 17. **June 2003 administration (*N* = 59).**

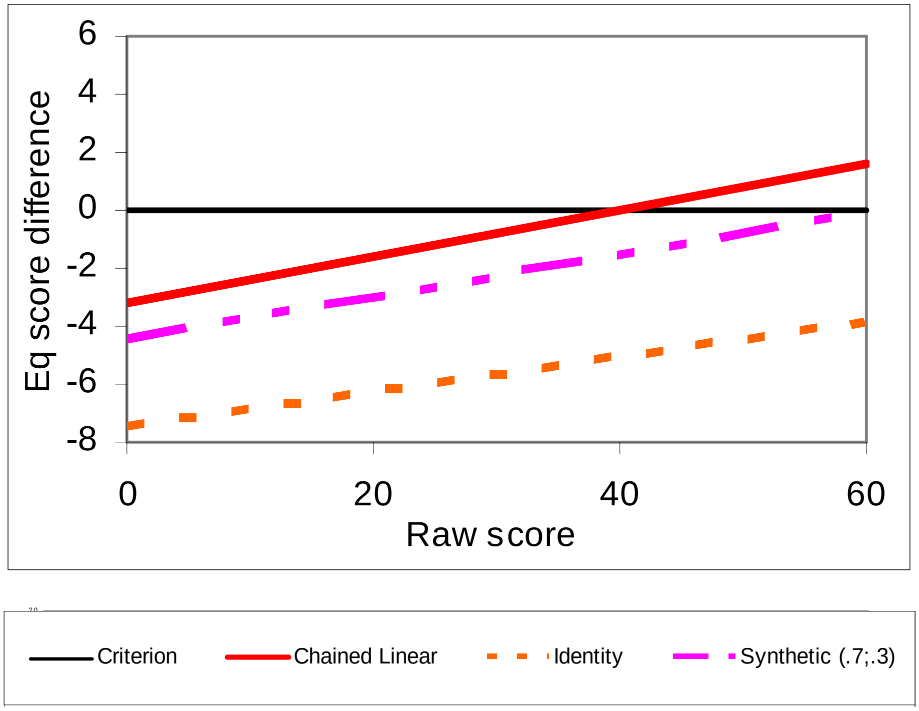


***Figure 18*. June 2005 administration (*N* = 65).**

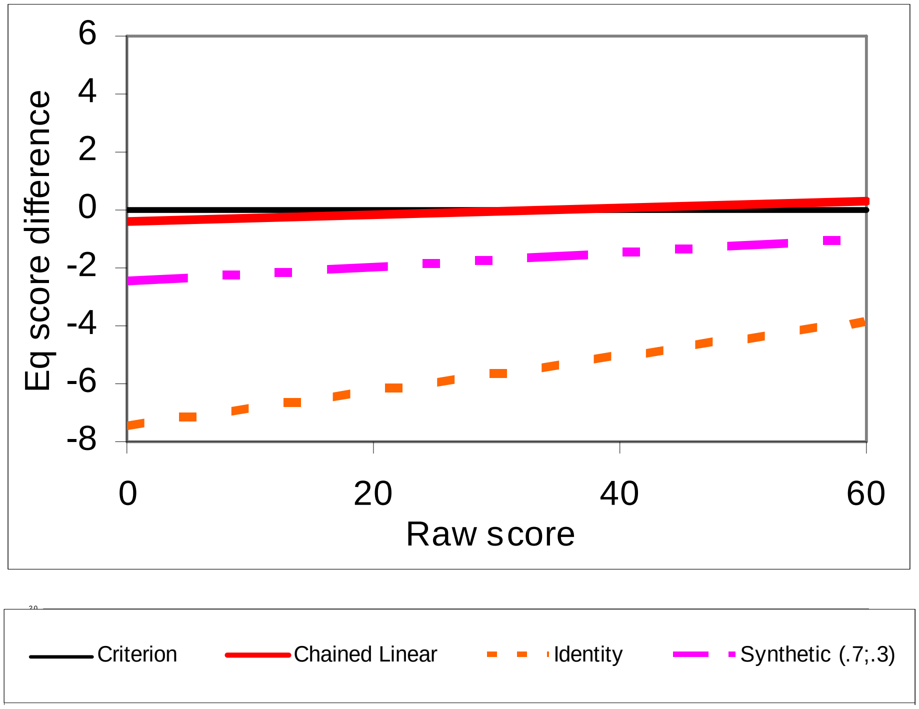


***Figure 19*. June 2004 administration (*N* = 69).**

similar, yielding the same pass/fail designations for the examinees. Among the 18 administrations, the Tucker method showed the smallest RMSDs at 8 administrations. Mean, chained linear, and Levine methods showed the smallest RMSDs at 4, 4, and 3 administrations, respectively. This trend was no longer true when the RMSDs were calculated only for the cut-score region. Again, the identity function yielded the largest RMSD (3.38) for this region. When the four linear methods, except the synthetic functions, were compared, mean equating yielded the smallest RMSDs for this region over the 10 administrations, and chained linear, Tucker, and Levine methods yielded the smallest RMSDs at 2, 3, and 3 administrations, respectively.

The results seemed to vary based upon specific data characteristics, such as the mean and variance of the total and anchor scores. Although the linear equating methods yielded slightly different RMSDs, the values were always within the 90% CI of the smallest RMSD in each administration; this was true in both the entire score and cut-score ranges. This indicated negligible differences among the linear equating methods in dealing with small samples in test equating.

As mentioned, forms *X* and *Y* differed markedly in difficulty; thus, any form of equating seems to be better than no equating, even when the equating sample was as small as 25. The benefits of the synthetic functions, as our previous study indicated (Kim et al., 2006), were not clear for nonparallel test forms. In many administrations, the RMSDs of the synthetic function were out of the 90% CI of the smallest RMSD, which was calculated from one of the linear equating methods.

In general, the effectiveness of the synthetic function was questionable for nonparallel test forms. However, an interesting finding emerged for the extremely small samples. Although the two forms vary widely in difficulty, the use of the identity function (i.e., no equating) seems to have some benefits when equating samples are extremely small ($N < 25$). The synthetic functions that used unequal weights to combine chained linear (.7) with identity (.3) showed the smallest RMSDs. As presented in Figures 2 to 4, the synthetic function performed well-compared to the mean and chained linear functions. The use of synthetic function will be very limited unless test forms are very similar in difficulty.

Study 2

Data

Although more examinees were available in Study 2 than in Study 1, the number was still less than 100 in each administration. The descriptive information for the data sets used in Study 2 is presented in Table 7. As shown, the data sets for form *X* were composed of relatively small samples ranging from 31 to 70. Each data set consisted of the raw sample frequencies of scores for two nonparallel, 119-item tests[4] with 46 internal anchor items given to two samples (*P* and *Q*) from a national population of examinees. Sample *P* included examinees who took the new test form *X*, and sample *Q* included those who took the reference form *Y*. As summarized in Table 8, the total number of examinees who took form *X* from April 2004 to November 2005 was 319,[5] and the total number who took form *Y* from November 2002 to January 2006 was 810. Descriptive statistics for these groups are summarized in Table 8.

Table 7

Descriptive Statistics of Each Sample in Study 2

Test form	Administration		*N*	Total		Anchor		r_{xv}
				M	*SD*	*M*	*SD*	
Form *X*	August	2005	31	94.23	11.09	38.52	4.36	.91
	September	2004	35	84.49	12.15	34.54	5.28	.91
	January	2005	56	92.05	12.39	37.59	4.55	.92
	April	2004	64	88.67	12.57	36.28	5.30	.93
	November	2005	66	87.86	13.25	35.80	5.17	.92
	April	2005	70	89.64	10.98	36.70	4.51	.91
Form *Y*			810	91.20	12.28	36.55	5.52	.92

Note. r_{xv} indicates correlation between total and anchor scores.

As shown in Table 8, the mean of the anchor test *V* was 36.53 (±0.28) in total group *P*, and 36.55 (± 0.19) in *Q*, where 0.28 and 0.19 were the standard errors of the mean. Thus, total group *P* was as proficient as total group *Q*, as measured by *V*. No ability difference existed between *P* and *Q*, although score variability somewhat differed. The psychometric properties were fairly similar for the two forms. The internal consistencies of the anchors were lower than those of the total tests for the same reason as in Study 1. The correlations between tests and anchors was reasonably high (r = .92).

Table 8

Summary Statistics for the Observed Distributions of X, V in P and Y, V in Q: Study 2

	N	μ	σ	SEM	Reliability	ρ
X_P	319	89.45	12.37	4.29	.88	.92
V_P		36.53	4.98	2.54	.74	
Y_Q	810	91.20	12.28	4.07	.89	.92
V_Q		36.55	5.52	2.53	.79	

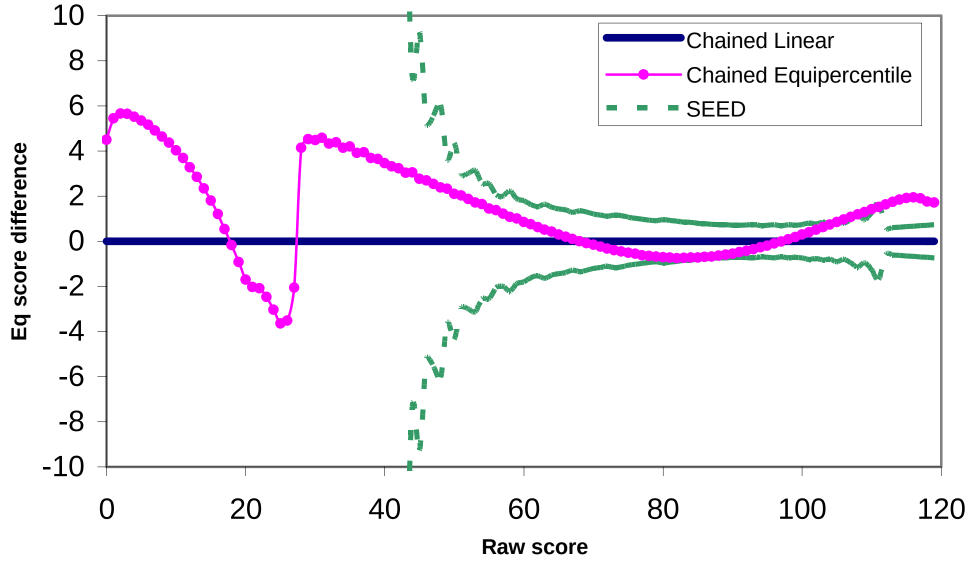
Note. SEM = Standard error of measurement. ρ = Correlation between total score and anchor. *P* = Accumulated from April 2004 to November 2005 administrations. *Q* = Accumulated 15 administration which were held from November 2002 to January 2006.

Criterion Function

Equating was conducted with a total of 319 examinees for form *X* and 810 examinees for form *Y*. For each raw score on form *X*, the equivalent raw scores on form *Y* were determined using the chained linear, Tucker, Levine, and chained equipercentile methods. As shown in Figure 20, the differences between the chained linear and chained equipercentile functions were very large for the score points from 0 to 60 and from 105 to 119. However, the differences were within the error band representing plus or minus two empirical conditional standard error of equating difference (SEED), and more importantly, almost no data fell in that region. Among the 319 examinees who took the new form *X*, the minimum score was 49, and only 3% (N = 8) of examinees received scores lower than 60. Again, the differences observed at this region were not substantial; thus, linear equating was selected as a criterion.

Results

The same procedures and weight systems used in Study 1 were applied to Study 2 based on the same rationale. We obtained the form *X* scores equated to form *Y* with six different samples (see Table 7), respectively, using identity, mean, chained linear, Levine, Tucker, and various synthetic function methods. The total examinees (called *Q*) who took form *Y* from



Raw Score X	Frequency
0-5	0
6-10	0
11-15	0
16-20	0
21-25	0
26-30	0
31-35	0
36-40	0
41-45	0
46-50	1
51-55	0
56-60	7
61-65	5
66-70	15
71-75	19
76-80	26
81-85	39
86-90	38
91-95	51
96-100	56
101-105	42
106-110	17
111-115	3
116-120	0
Total	319

***Figure 20.* Difference plot, chained linear versus chained equipercentile, and frequency distribution of form *X* scores in total group *P* in Study 2.**

Table 9

RMSD Between the Criterion Function and Sample-Based Linking Functions (Identity, Mean, Chained Linear, and Synthetic) Across the Entire Score Region: Study 2

N	Identity		Mean		Chained linear		Synthetic (.5CH+.5ID)		Synthetic (.7CH+.3ID)	
31	2.14	(1.96)	1.30	(1.01)	0.36	(0.34)	0.85	(0.77)	0.36	(0.32)
35	2.14	(1.96)	1.38	(0.76)	0.96	(0.49)	1.27	(1.19)	1.03	(0.89)
56	2.14	(1.96)	1.30	(0.98)	0.99	(0.86)	0.70	(0.50)	0.45	(0.14)
64	2.14	(1.96)	1.31	(0.88)	0.56	(0.31)	1.17	(1.11)	0.84	(0.78)
66	2.14	(1.96)	1.31	(1.07)	0.41	(0.16)	1.05	(0.89)	0.67	(0.46)
70	2.14	(1.96)	1.31	(0.88)	0.31	(0.09)	1.04	(0.98)	0.64	(0.60)

Note. Cut-score ranges in parentheses.

Table 10

RMSD Between the Criterion Function and Sample-Based Linking Functions (Tucker, Levine, and Synthetic) Across the Entire Score Region: Study 2

N	Tucker		Synthetic (.5Tucker +.5ID)		Synthetic (.7Tucker +.3ID)		Levine		Synthetic (.5Levine +.5ID)		Synthetic (.7Levine +.3ID)	
31	0.28	(0.26)	1.18	(1.09)	0.82	(0.76)	0.62	(0.58)	0.71	(0.64)	0.18	(0.14)
35	1.24	(0.42)	1.22	(1.11)	1.06	(0.79)	0.86	(0.53)	1.29	(1.22)	1.04	(0.93)
56	0.62	(0.50)	0.85	(0.70)	0.45	(0.23)	1.13	(1.00)	0.64	(0.43)	0.49	(0.21)
64	0.64	(0.35)	1.19	(1.13)	0.88	(0.80)	0.53	(0.30)	1.17	(1.10)	0.83	(0.77)
66	0.25	(0.12)	1.02	(0.90)	0.60	(0.49)	0.49	(0.17)	1.06	(0.89)	0.70	(0.49)
70	0.66	(0.40)	1.22	(1.15)	0.91	(0.84)	0.23	(0.09)	0.98	(0.92)	0.54	(0.51)

Note. Cut-score ranges in parentheses.

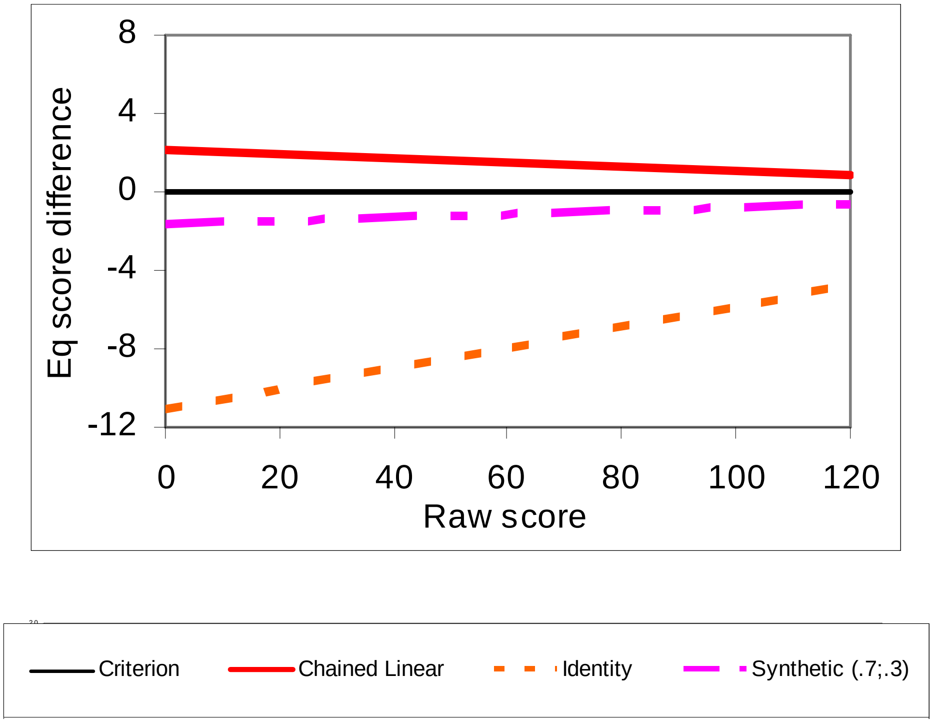


***Figure 21*. August 2005 administration (*N* = 31).**

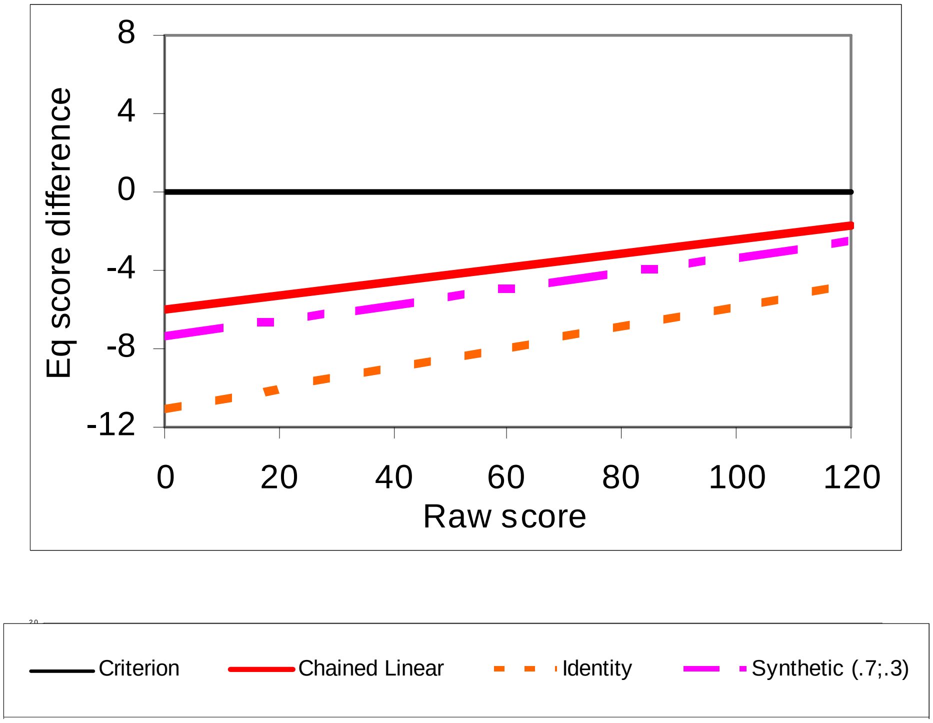


***Figure 22*. September 2004 administration (*N* = 35).**

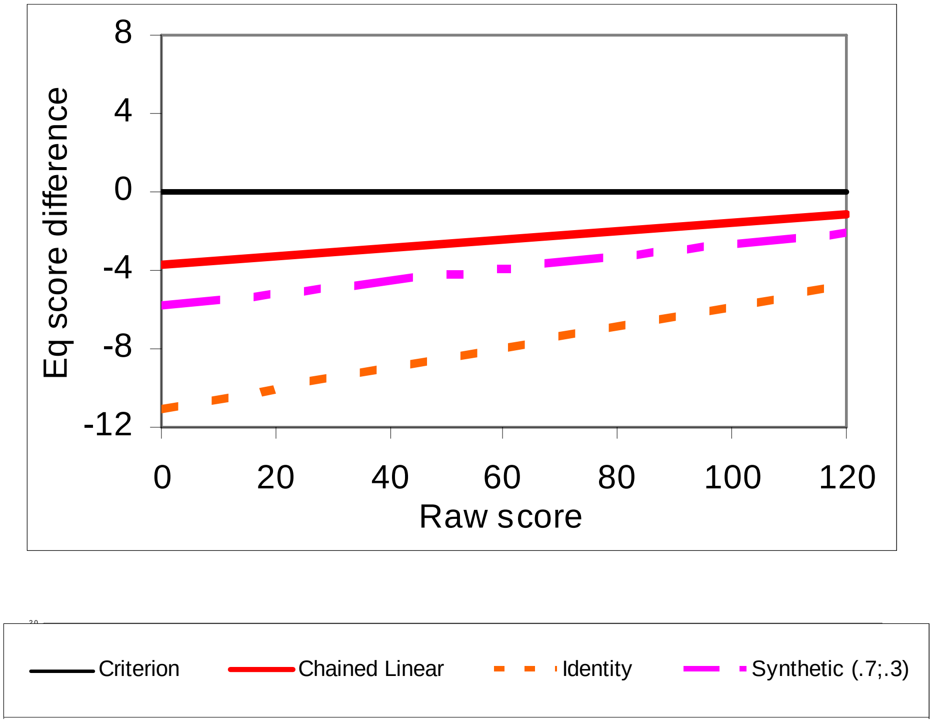


***Figure 23*. January 2005 administration (*N* = 56).**

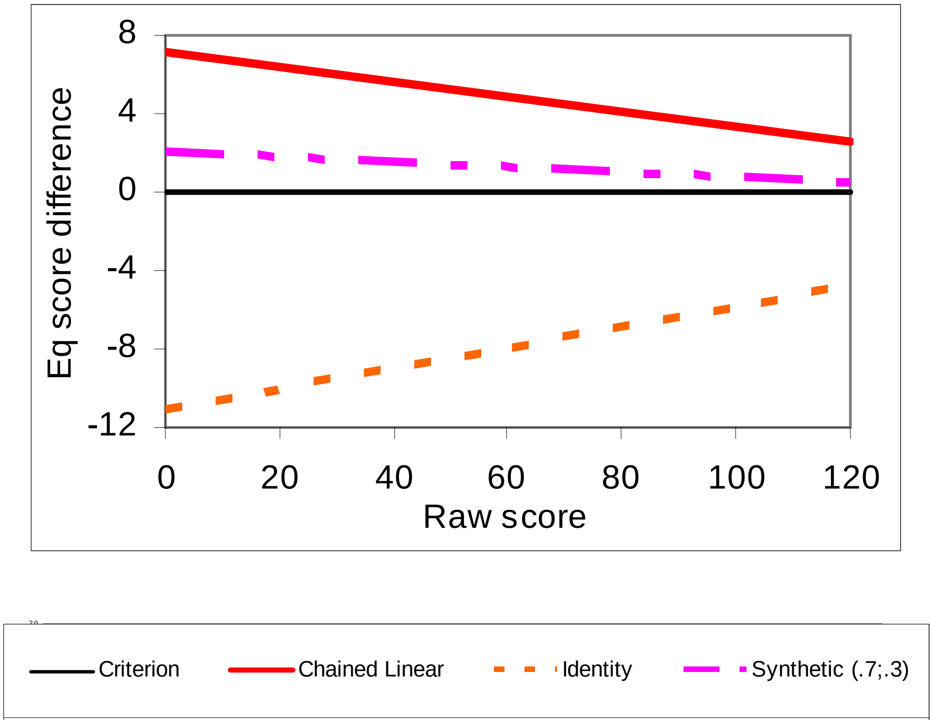


***Figure 24*. April 2004 administration (*N* = 64).**

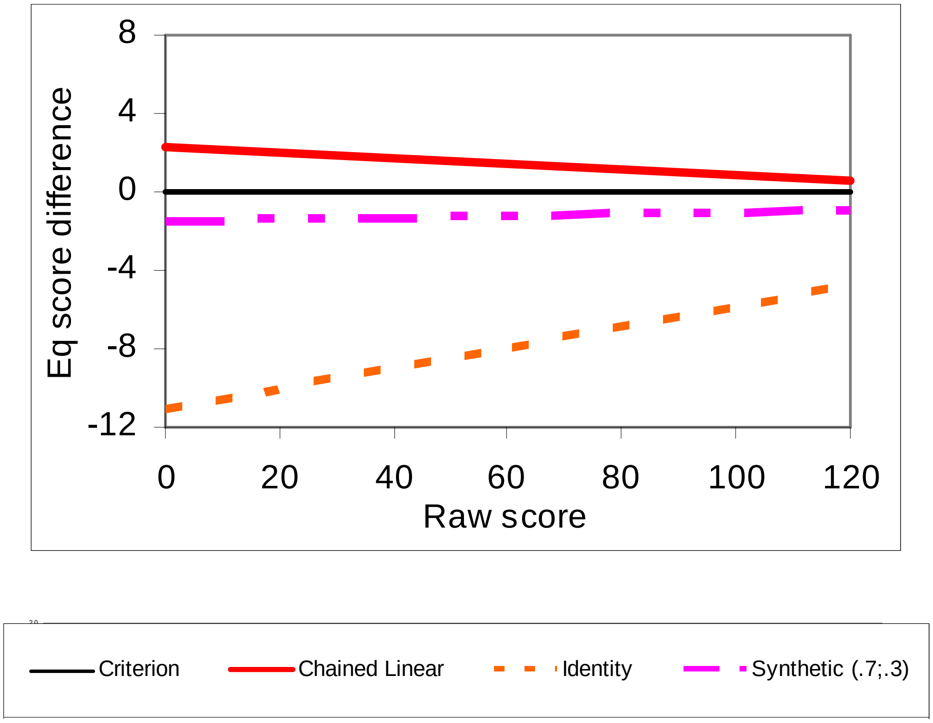


***Figure 25*. November 2005 administration (*N* = 66).**

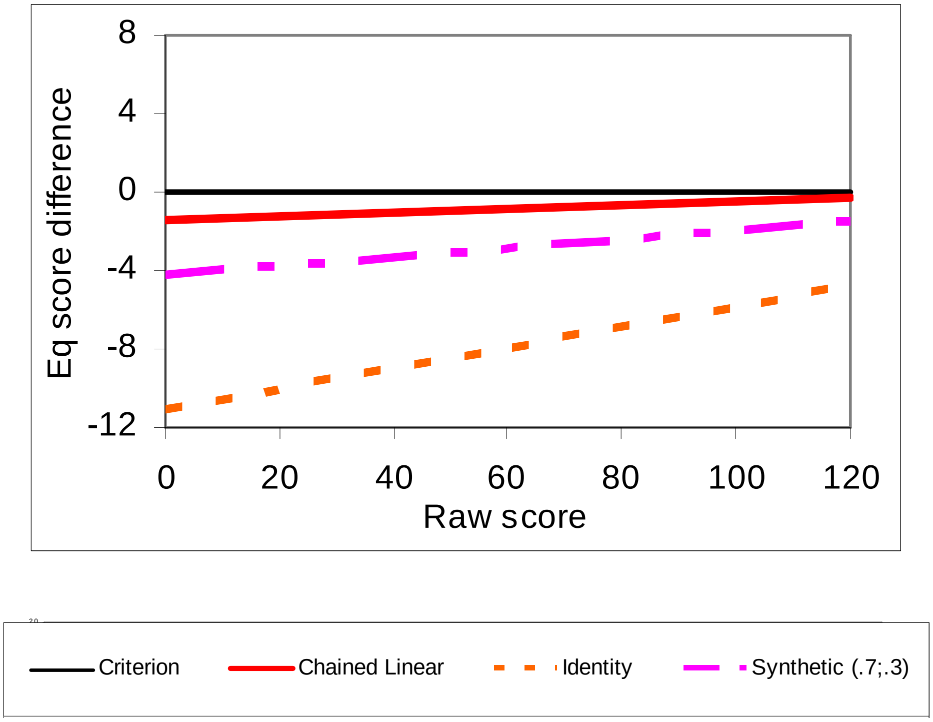


***Figure 26*. April 2005 administration (*N* = 70).**

November 2002 to January 2006 (*N* = 816) were used as the reference sample in each administration. The differences between each sample equating and the criterion were calculated with respect to the RMSD deviance measure.

Tables 9 and 10 present RMSDs across the entire range of raw scores and the cut-score region (Raw Scores 50 to 87). Figures 21 to 26 plot the difference between criterion and each equating/linking function (identity, chained linear, and a synthetic function). As in Study 1, two weight systems (e.g., equal weights and more weights on equating function) were used to synthesize two different functions. With respect to RMSDs, the general trends of linking/equating functions were quite similar for both the entire score and cut-score regions.As summarized in Table 8, examinees (called *P*) who took form *X* were as proficient as examinees (called *Q*) who took form *Y*, as measured by the anchor. Despite equal ability in both groups, the mean of form *X* (*M* = 89.45) was lower than the mean of form *Y* (*M* = 91.20). In terms of effect sizes, the difference between these two means (-1.75) is approximately 14% of the averaged standard deviation of 12.33. The new form *X* was somewhat harder than the reference form *Y*. Although the difference between forms here was not as large as in Study 1, its standardized mean difference (0.14) was fairly large. Because forms *X* and *Y* were not parallel, as in Study 1 the identity function yielded considerable bias in Study 2. As presented in Tables 9 and 10, the RMSDs of the identity function were 2.14 over the entire score range and 1.96 over the cut-score region. As a result, the synthetic functions also yielded relatively large RMSDs compared to the other linear equating methods.

As in Study 1, three linear methods (except the mean equating) worked similarly over six administrations. The Tucker and Levine methods yielded the smallest RMSDs at the three administrations, respectively (see Tables 9 and 10); however, RMSDs derived from other linear methods were within the 90% CI of the smallest RMSD in each administration, as shown in Tables 11 and 12. As expected, no clear evidence supported a certain method in small sample equating situations. Interestingly, mean equating did not work as well as other linear methods in this case. Its RMSDs were quite large regardless of sample size. Mean equating assumes differences in difficulty constant throughout score ranges, but this assumption might not be true in this case, as shown in Figure 20. form *X* might be more difficult than form *Y* for more able examinees. Equating methods that allowed the relative difficulty of the forms to vary along the score range seemed to be more useful in this case.

Table 11

The 90% Confidence Interval for RMSDs Calculated for the 100 Bootstrap Samples (Mean, Chained Linear, and Synthetic) Across the Entire Score Region: Study 2

N	Mean	Chained linear	Synthetic (.5CH+.5ID)	Synthetic (.7CH+.3ID)
31	(1.30, 2.35)	(0.27, 2.28)	(0.16, 2.16)	(0.17, 2.17)
35	(1.20, 2.43)	(0.28, 2.75)	(0.79, 2.11)	(0.58, 2.31)
56	(1.30, 1.77)	(0.23, 1.86)	(0.31, 1.23)	(0.18, 1.18)
64	(1.30, 1.78)	(0.25, 1.53)	(0.83, 1.70)	(0.44, 1.59)
66	(1.30, 1.68)	(0.13, 1.30)	(0.69, 1.46)	(0.20, 1.30)
70	(1.30, 1.74)	(0.24, 1.59)	(0.73, 1.44)	(0.27, 1.35)

Table 12

The 90% Confidence Interval for RMSDs Calculated for the 100 Bootstrap Samples Across the Entire Score Region (Tucker, Levine, and Synthetic): Study 2

N	Tucker	Synthetic (.5Tucker +.5ID)	Synthetic (.7Tucker +.3ID)	Levine	Synthetic (.5Levine +.5ID)	Synthetic (.7Levine +.3ID)
31	(0.35, 2.98)	(0.37, 2.55)	(0.28, 2.72)	(0.29, 2.44)	(0.13, 2.03)	(0.10, 2.00)
35	(0.42, 3.20)	(0.87, 2.17)	(0.49, 2.50)	(0.25, 2.72)	(0.72, 2.08)	(0.54, 2.19)
56	(0.25, 1.47)	(0.39, 1.45)	(0.19, 1.24)	(0.40, 2.09)	(0.23, 1.18)	(0.20, 1.18)
64	(0.27, 1.78)	(0.84, 1.78)	(0.51, 1.69)	(0.26, 1.47)	(0.82, 1.69)	(0.44, 1.58)
66	(0.15, 1.13)	(0.68, 1.45)	(0.23, 1.23)	(0.19, 1.44)	(0.70, 1.49)	(0.22, 1.36)
70	(0.26, 2.01)	(0.86, 1.69)	(0.44, 1.69)	(0.20, 1.59)	(0.68, 1.41)	(0.23, 1.27)

The chained linear method adjusts for differences in group ability using the anchor test, then adjusts for test difficulty based on the adjusted test score means and standard deviations. If the difference of the form *X* mean from the form *X* criterion group mean was proportional to the difference of the anchor mean from the anchor criterion group mean (assuming equal *SD*s) for every administration, all conversions would be the same with the criterion. When the administration form *X* mean is lower than expected given the anchor test mean (when compared with the criterion group), a higher conversion for the administration than for the criterion will emerge. For the January, August, and November 2005 administrations, sample *P*'s performance on the anchor test suggested a higher performance on form *X* than was actually observed, judged by the performance of the criterion group on the anchor and form *X* (see Table 8). For these three administrations, the form *X* conversion is higher than in the criterion group, as presented in Figures 21, 23, and 25. Consequently, the synthetic function worked better than did the other functions, particularly for August 2005 ($N = 31$) and January 2005 ($N = 56$). The synthetic functions counterbalanced those functions, identity and any format of equating, showing the smallest RMSD over the entire score range, including the cut-score region.

Discussion

When equating it is desirable to have a large enough sample to produce stable and accurate results. Many testing programs (e.g., certification tests), however, are low volume in nature; therefore, it is often hard to obtain as many as 50 examinees for test equating. In a previous study (Kim et al., 2006), we introduced an approach, called the synthetic linking function, to conduct test linking with small samples. Essentially, the synthetic function is a compromise between the identity function and sample equating. In the previous study, the synthetic function provided certain empirical benefits, including reduction of bias and linking error, with small samples randomly selected from large operational samples. It was concluded that the synthetic function method might be an alternative when sample sizes are small and groups differ in ability in situation where two forms are well-designed and almost parallel.

We conducted the present study to extend our previous work by investigation of the synthetic function in situation where testing conditions are less attractive. In the present study, operational data sets of two low-volume subject tests from a licensure program were used to examine the effectiveness of the synthetic function. The test forms used here were clearly nonparallel. We compared the linking results of the identity, synthetic, chained linear, mean, Tucker, and Levine functions with the linking criterion with respect to the RMSD index. As mentioned previously, we chose linear equating derived from the total samples as the equating criterion in both studies. It may not be apparent why the total sample criterion function is linear in nature. Because the actual samples of both tests used in this study were very small, their criteria were not derived from substantially large samples (less than 1,000). Although the results based on the linear criterion are presented in this paper, we also calculated the RMSD index using the nonlinear (chained equipercentile) criterion. The actual RMSD numbers differed slightly, but the trends and conclusions were identical.

In Study 1, the synthetic function did a better job than the traditional linear functions when the sample size was very small (less than 30), consistent with our previous investigation (Kim et al., 2006). With very small equating samples, the synthetic function produced a smaller linking error than did other traditional methods and provided less bias than the identity function did. The synthetic function seems to be a better choice than the sample equating function in some contexts. With samples larger than 30, however, some form of equating was clearly preferable to no equating. Because forms *X* and *Y* were clearly not parallel, the identity function yielded the largest bias. Similar results emerged from Study 2, indicating limitations of the identity function. However, it is worth noting that sample equating functions (e.g., chained linear, Tucker, or Levine) also showed substantial RMSDs for small samples ($N < 40$), implying incorrectly specified pass/fail designations for examinees. Although RMSDs tend to decrease as the sample size increases, equating with very small samples remains problematic.

Whether or not to use the identity function is a major decision when dealing with small samples. As mentioned, this decision may depend on several variables, such as specific sample size, location of passing scores on the raw score scale (if applicable), equating design (e.g., NEAT or random group design), and degree of difference between forms. For example, Skaggs (2005) recommended the use of the identity function in the random group case when forms differ by one-tenth of a standard deviation or less. Using the identity function may allow a small amount of equating error when test forms are carefully developed from the same set of test specifications. Ironically, well-controlled test assembly for a stable test is usually accompanied by ample data for equating. A lack of data is likely to affect the test assembly process as well as the equating process. This means that the use of the identity function instead of some form of equating may lead to a large amount of unknown bias, called systematic error.

In general, the total equating error can be partitioned into random error and systematic error components. Which is more worrisome in small sample equating situations: bias or equating error (i.e., random error) resulting from sample variability? The answer may depend on the situation. For example, bias may be especially problematic for tests with cut scores. The sample equating function may have less linking bias than the identity function; however, for small samples, it has quite a bit of error due to sample variability. Conversely, the identity is usually quite biased even though it has no sample variability. The sum of random equating error variance and squared bias equals the mean squared error in equating. The intent in using the identity function is for the increase in systematic error to be more than offset by the decrease in random error. In practice, the decision whether or not the identity function is preferable to other equating functions may be sensitive to the extent that the forms are assumed to differ, along with the degree of difference between distributions of the scores and the format of equating designs (see Kolen & Brennan, 2004, p. 289).

This study was designed to investigate the effectiveness of the synthetic linking function, which is a weighted average of the identity function (having no equating error but large bias) and the traditional equating function (having small bias but large equating error), using data sets from operational administrations. As summarized, the benefits of the synthetic function were limited by pronounced differences in form difficulty. In many testing programs, test forms are designed to be parallel. Given that assumption, the synthetic function offers a useful compromise between the two and can better reduce the total mean squared error than can the sample equating function or the identity function when equating samples are extremely small. For that reason, the synthetic function might be an alternative where test equating must be based on very small samples; however, due to some practical limitations (e.g., no pretested items) this is not always the case, as proven in the present study. Again, it is worth noting that we do not provide the synthetic function as a solution or methodological fix to a problem that is caused by poor data collection practice.

Further research is necessary to discover the proper development of weights to use when synthesizing the equating function with the identity function. It is not clear how to appropriately weight the identity and conventional equating functions in the absence of either historical information concerning variability of form difficulty or specific information concerning form construction. An objective tool to guide weighting is unavailable. One possibility is to collect empirical information from previous administrations or different forms of the same test. A simulation study designed to establish a procedure for defining a priori information using historical test information is ongoing.

References

Dorans, N. J., & Feigenbaum, M .D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10, pp. 91-122). Princeton, NJ: ETS.

Harris, D. J. (1993, April). *Practical issues in equating.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Kim, S., von Davier, A. A., & Haberman, S. (2006, April). *An alternative to equating with small samples in the non-equivalent groups anchor test design*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer Science & Business Media.

Livingston, S. A. (1993). Small sample equating with long-linear smoothing. *Journal of Educational Measurement, 30*, 23–39.

Parshall, C. G., Du Bose Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*, 37–54.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*, 309–330.

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the non-equivalent group design. *Journal of Educational and Behavioral Statistics, 30*, 313–342.

Notes

[1] It is worth noting that all states do not use the same cut score.

[2] Although one item (Item 13) was not scored in form *X* due to an ambiguity in the stem, this item was scored here so that the possible raw score range of the test forms (*X* and *Y*) was in the same order, so that the identity function can be used later in a simple manner.

[3] The total number of examinees was smaller than the total number of examinees from each of the individual administrations, because only the first record was included for test repeaters.

[4] Due to ambiguity in item content, one item was not scored in each test form, *X* and *Y*. As a result, the possible raw score range of the test forms (*X* and *Y*) was the same (119, not 120).

[5] The total number of examinees was smaller than the total number of examinees from each of the individual administrations because only the first record is included for test repeaters.